Towards Robust Identification of Rail Short-Wavelength Irregularities: A Multi-Modal Fusion of Vibration and Profile Data

Yiming Zhai, Huanyu Yang, Jun Wang, and Yuntong An

***Abstract*—Short-wavelength rail irregularities (SWRIs) are localized geometric deviations that evolve under periodic wheel loads and can compromise operational safety. Vibration-only inspection is prone to noise and response ambiguity, whereas profile-only inspection lacks temporal dynamics of wheel–rail interaction. We propose a multi-modal framework that fuses axle-box vibration sequences with high-resolution rail profile data. On the vibration side, features are learned via stacked autoencoders (SAE); on the profile side, measured contours are registered to a standard template, transformed into wear sequences via Dynamic Time Warping (DTW), and compressed with Principal Component Analysis (PCA). Two kernel SVMs (KSVMs) trained on the respective modalities are aggregated through a class-dependent confidence fusion. On a 100 m laboratory rail with simulated grinding (F1), spalling (F2), abrasion (F3), and normal (F4), we collect 200 paired samples and show that the fusion model markedly reduces misclassification, achieving accuracies of 87.5% (F1), 100% (F2), 93.3% (F3), and 94.1% (F4), with overall gains ranging from 7.1% to 47.7% over single-source baselines. The approach provides a principled path to robust, intelligent rail condition monitoring.**

***Key Words*—Railway safety, short-wavelength irregularity, vibration analysis, rail profile, multi-modal fusion**

I. Introduction

ENSURING the structural integrity of railway infrastructure has long been a cornerstone of safe and efficient transportation systems [1]. Among the various degradation phenomena that occur on rails, short-wavelength irregularities represent one of the most challenging threats. These defects, typically characterized by localized geometric deviations with periodic patterns, emerge gradually under the repeated action of wheel loads and can severely deteriorate ride comfort, accelerate rail wear, and, in extreme cases, precipitate derailment accidents [2]. As railway networks expand and traffic intensity continues to rise, the early detection of such irregularities becomes an increasingly critical requirement for both operational safety and cost-effective maintenance [3].

Traditional approaches to track condition monitoring have largely relied on vibration-based sensing. By attaching accelerometers to the axle-box, the dynamic response of the wheel–rail interaction can be continuously monitored. This technique offers attractive advantages, such as low cost and real-time capability [4], but it often suffers from signal contamination by ambient noise and ambiguous mapping between vibration patterns and defect categories [5]. Different irregularity types or depths may yield overlapping vibration signatures, thereby reducing diagnostic reliability.

As technology progresses, machine vision has been widely applied across fields such as military, mining, medicine, and agriculture [5–8], and is increasingly becoming an important tool for rail defect detection. Machine vision works by converting image information into digital data for analysis and recognition, using high-resolution imaging devices (e.g., cameras or drones) to capture photographs of railway tracks, followed by image processing and computer-vision algorithms to identify damage [9]. Although this approach offers high detection accuracy and good continuity, images are vulnerable to noise, occlusion, or low contrast, and visual imaging alone cannot comprehensively assess rail geometry or the depth distribution of short-wavelength irregularities. Consequently, the effectiveness of machine vision under complex operating conditions remains limited. Parallel efforts have introduced optical and profile measurement technologies [10], which provide a more direct view of the rail geometry. Laser-based contour acquisition, for example, generates high-resolution two-dimensional point clouds of the rail cross-section [11], allowing wear and deformation to be quantified with high accuracy [12, 13, 14]. However, profile inspection on its own fails to capture the dynamic interaction effects that occur when trains are in motion, and its ability to characterize defect evolution across multiple temporal scales is limited.

In light of these challenges, data fusion emerges as a promising avenue. By combining the complementary strengths of vibration and geometric profile measurements, a more holistic description of rail conditions can be achieved. This paper proposes a multi-modal fusion framework that leverages deep representation learning and statistical modeling to jointly interpret both data modalities, ultimately aiming to enhance the robustness and precision of short-wavelength irregularity detection.

Contributions. We present a decision-level multi-modal framework that (i) learns compact vibration representations via SAE; (ii) converts registered contours into temporal wear sequences with DTW and reduces them by PCA; (iii) trains per-modality KSVMs; and (iv) fuses probabilistic outputs with *class-dependent* reliability weights. On a controlled 100 m

Manuscript received September 1, 2025; revised September 8 and September 15, 2025; accepted October 7, 2025. This article was recommended for publication by Associate Editor Shujin Qin upon evaluation of the reviewers' comments.

Y. Zhai, H. Yang, and J. Wang are with the School of Automation, Nanjing University of Science and Technology, Nanjing 210000, China (e-mail: zhaiyiming@njust.edu.cn; yhy@njust.edu.cn; wangj1125@163.com).

Y. An is with the College of Automation and Intelligence, Beijing Jiaotong University, Beijing 10000, China (e-mail: 21111059@bjtu.edu.cn).

Corresponding author: Jun Wang

testbed with four classes (F1–F4), we verify consistent gains of 7.1%–47.7% over single-modality baselines, and analyze why fusion reduces confusions between grinding (F1) and abrasion (F3).

II. Related Work

A. Vibration-based rail monitoring

Vibration analysis has been one of the earliest and most widely used methods for track condition monitoring. Accelerometers mounted on bogies or axle-boxes record the vertical and lateral dynamic responses induced by wheel–rail contact [12]. Several studies have demonstrated the feasibility of identifying corrugation, spalling, and similar defects from spectral features of vibration signals. Nevertheless, ambiguities arise because different defect categories can yield comparable vibration signatures under varying operational conditions, such as speed or axle load. Moreover, background vibrations from the vehicle body and surrounding environment often obscure subtle irregularity patterns, leading to false positives and inconsistent recognition accuracy. [2, 4]

B. Optical vision and rail profile metrology

With advances in computer vision, image-based approaches have been introduced for rail defect detection. High-speed cameras and unmanned aerial vehicles (UAVs) enable continuous visual inspection, followed by digital image processing to highlight cracks, scratches, or material loss. These methods have proven effective in capturing visible surface anomalies and offer high spatial resolution [12]. However, their effectiveness is heavily dependent on environmental conditions such as illumination, occlusion, or dirt on the rail surface. Furthermore, purely visual systems struggle to assess the depth and periodicity of short-wavelength irregularities, limiting their applicability to dynamic safety assessment. [10, 11, 15]

Profile-based detection addresses part of this limitation by directly quantifying rail cross-section geometry. Laser profilers project structured light onto the rail and reconstruct contour data as two-dimensional point clouds [13]. When compared against a reference rail template, deviations in head and waist regions can be identified as indicators of wear and deformation [14]. While profile measurement has been widely adopted in heavy-haul railways for wear monitoring, it alone does not capture the temporal dynamics associated with vibration responses, making it insufficient for a complete diagnosis of short-wavelength defects.

C. Multi-source fusion in infrastructure

To overcome the drawbacks of single-modality approaches, multi-source data fusion has gained momentum in infrastructure health monitoring. Applications in bridge inspection, rotating machinery diagnostics, and pavement evaluation have demonstrated the benefits of integrating complementary data types to improve reliability and resilience of detection systems. In the railway domain, however, fusion studies remain relatively sparse. Existing research has mainly focused on combining multiple vibration channels or integrating visual data with limited sensor inputs. Comprehensive frameworks that combine both vibration dynamics and rail profile information are still lacking, particularly for short-wavelength irregularity detection. [6, 7, 8, 9]

III. Problem Formulation and Notation

Motivated by these gaps, we now cast short-wavelength irregularity detection as a supervised classification problem. We first specify the paired sensing dataset, the modality-specific embeddings, and the probabilistic decision rule adopted in this work. Let $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{p}^{(n)}, y^{(n)})\}_{n=1}^{N}$ be paired Axle Box Acceleration (ABA) sequences $\mathbf{x}^{(n)} \in \mathbb{R}^{T_n}$, profile samples $\mathbf{p}^{(n)} \in \mathbb{R}^{M\times 2}$ (2D points), and labels $y^{(n)} \in \{1, 2, 3, 4\}$ (F1: grinding, F2: spalling, F3: abrasion, F4: normal). In practice, T_n is the time length (i.e., the vector length) of the vibration signal for sample n, which may vary across samples due to factors such as running speed and window length, while M denotes the number of valid contour points after masking and ordering. We assume that $(\mathbf{x}^{(n)}, \mathbf{p}^{(n)})$ are synchronized at the segment level, i.e., they correspond to the same track location within a small spatial tolerance, which enables effective cross-modality reasoning.

Our goal is a classifier $f : \mathbb{R}^{d_v} \times \mathbb{R}^{d_g} \to \{1..4\}$ with $d_v = 20$ (vibration embedding) and $d_g = 11$ (geometric embedding). Concretely, we consider modality-specific mappings

$$\phi_v : \mathbb{R}^{T_n} \to \mathbb{R}^{d_v}, \qquad \phi_g : \mathbb{R}^{M\times 2} \to \mathbb{R}^{d_g},$$

which yield $\mathbf{v}^{(n)} = \phi_v(\mathbf{x}^{(n)})$ and $\mathbf{g}^{(n)} = \phi_g(\mathbf{p}^{(n)})$. The classifier then acts on $(\mathbf{v}^{(n)}, \mathbf{g}^{(n)})$ to produce a discrete decision $\hat{y}^{(n)} = f(\mathbf{v}^{(n)}, \mathbf{g}^{(n)})$. We further write the classifier in probabilistic form via calibrated posteriors $\pi_k(\mathbf{v}, \mathbf{g}) = \mathbb{P}(y = k \mid \mathbf{v}, \mathbf{g})$, with the Bayes decision $\hat{y} = \arg\max_k \pi_k(\mathbf{v}, \mathbf{g})$. To avoid ambiguity, we denote class priors by $\rho_k = \mathbb{P}(y = k), \quad k \in \{1, \ldots, 4\}$.. This formulation allows us to incorporate class-dependent costs or abstention rules if needed.

A Short Wavelength Rail Irregularity (SWRI) is a localized periodic deviation with spatial wavelength $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. If $z(s)$ is the longitudinal rail surface profile along arc-length s, the SWRI set can be modeled as

$$\mathcal{S} = \left\{ z : \ \exists \lambda \in \Lambda, \ \exists s_0, \ \|z(s) - \bar{z}(s)\|_{\mathcal{H}} \geq \tau, \ s \in [s_0, s_0 + \lambda] \right\}, \tag{1}$$

with $\bar{z}$ a nominal surface and $\|\cdot\|_{\mathcal{H}}$ a Sobolev seminorm encoding smoothness. Intuitively, the condition $\|z - \bar{z}\|_{\mathcal{H}} \geq \tau$ captures both amplitude and roughness of the deviation (quantification of the high-frequency energy or curvature degree of the deviation within a local region), filtering out benign, slowly varying wear while emphasizing short-wavelength, high-curvature defects.

For completeness, we make two mild measurement assumptions. First, the observed vibration and profile arise from latent, noise-free signals corrupted by zero-mean disturbances: $\mathbf{x}^{(n)} = \mathbf{x}_{\star}^{(n)} + \boldsymbol{\varepsilon}_v^{(n)}$ and $\mathbf{p}^{(n)} = \mathbf{p}_{\star}^{(n)} + \boldsymbol{\varepsilon}_g^{(n)}$, where $\boldsymbol{\varepsilon}_v^{(n)}$ models sensor noise and operational perturbations (e.g., unsprung mass dynamics), and $\boldsymbol{\varepsilon}_g^{(n)}$ reflects scanning noise and alignment errors, $\mathbf{x}_{\star}^{(n)} \in \mathbb{R}^{T_n}$ denotes the latent noise-free ABA sequence and $\mathbf{p}_{\star}^{(n)} \in \mathbb{R}^{M_n\times 2}$ the latent noise-free rail profile (2-D contour points in the local rail frame, before

alignment). The disturbances satisfy $\mathbb{E}[\boldsymbol{\varepsilon}_v^{(n)}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\varepsilon}_g^{(n)}] = \mathbf{0}$ and are assumed independent across modalities and samples. Second, the spatial–temporal link between dynamic response and wavelength is approximately $f_p \approx v/\lambda$ for train speed v and dominant vibration peak f_p, which rationalizes why speed normalization or window selection improves class separability.

Since profile points are collected in a local rail coordinate frame, we denote by $\mathcal{A}$ an alignment operator (rigid transform) that maps $\mathbf{p}^{(n)}$ to the reference template before feature extraction; formally, $\tilde{\mathbf{p}}^{(n)} = \mathcal{A}(\mathbf{p}^{(n)})$. The geometric mapping ϕ_g then acts on $\tilde{\mathbf{p}}^{(n)}$ (or on a derived wear sequence) to produce a compact descriptor. Likewise, ϕ_v may include normalization and segment selection to mitigate speed-induced spectral shifts.

Finally, to streamline notation in subsequent sections, we aggregate features as $\mathbf{z}^{(n)} = [\mathbf{v}^{(n)}; \mathbf{g}^{(n)}] \in \mathbb{R}^{d_v+d_g}$ and write class priors as $\pi_k = \mathbb{P}(y = k)$, $k \in \{1..4\}$. When discussing theoretical properties (e.g., calibration or risk bounds), we will refer to the conditional densities $p(\mathbf{v}, \mathbf{g} \mid y = k)$ and the induced decision rule $\hat{y}(\mathbf{v}, \mathbf{g})$, but the learning procedures themselves are entirely data-driven and do not assume parametric forms for these densities.

IV. Methodology

Fig. 1 summarizes the pipeline: ABA→SAE, profile→registration→DTW wear sequence→PCA, per-modality KSVMs, and class-dependent fusion. In other words, vibration signals are first preprocessed and embedded through a stacked autoencoder, while profile measurements are aligned, converted into wear sequences, and compressed via principal component analysis. Each modality is then classified independently using kernel SVMs, and their probabilistic outputs are integrated at the decision level. This layered architecture ensures that the strengths of both dynamic and geometric cues are preserved and that their weaknesses are mutually compensated.

A. Vibration preprocessing and segment selection

Raw sequences differ in length T_n because of variable speed, operating conditions, and sensor durations. To ensure comparability across samples, we first apply z-score normalization:

$$\tilde{x}_t^{(n)} = \frac{x_t^{(n)} - \mu_n}{\sigma_n}, \quad \mu_n = \frac{1}{T_n}\sum_{t=1}^{T_n} x_t^{(n)}, \ \sigma_n^2 = \frac{1}{T_n}\sum_t (x_t^{(n)} - \mu_n)^2. \tag{2}$$

This operation standardizes signals by removing mean offsets and scaling variances, thereby emphasizing informative fluctuations. Since not all parts of the sequence carry useful diagnostic information, we then select an informative window $\mathcal{W}_n$ that maximizes empirical variance:

$$\mathcal{W}_n = \arg\max_{W \subset \{1..T_n\}, |W|=L} \mathrm{Var}\big(\{\tilde{x}_t^{(n)} : t \in W\}\big). \tag{3}$$

Intuitively, windows with higher variance are more likely to capture defect-induced oscillations rather than background noise or steady-state vibration.

B. Stacked autoencoder (SAE) for vibration representation

We denote by $\mathbf{z}_0$ the variance-maximized segment extracted from $\mathbf{x}^{(n)}$. Let $\mathbf{z}_0 \in \mathbb{R}^L$ be the selected segment. To obtain compact and noise-robust features, we construct a two-layer SAE:

$$\mathbf{h}_1 = \phi(W_1\mathbf{z}_0 + \mathbf{b}_1), \quad \mathbf{z}_1 = \psi(\tilde{W}_1\mathbf{h}_1 + \tilde{\mathbf{b}}_1), \tag{4}$$

$$\mathbf{h}_2 = \phi(W_2\mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{z}_2 = \psi(\tilde{W}_2\mathbf{h}_2 + \tilde{\mathbf{b}}_2), \tag{5}$$

where ϕ is a nonlinear activation (ReLU or tanh) and ψ is linear. Hidden layer sizes are set to 100 and 20 to progressively compress the representation. The training objective balances reconstruction error and sparsity:

$$\mathcal{L}_{\mathrm{SAE}} = \frac{1}{L}\|\mathbf{z}_0 - \mathbf{z}_2\|_2^2 + \lambda\|\mathbf{h}_1\|_1 + \lambda\|\mathbf{h}_2\|_1. \tag{6}$$

Greedy layer-wise pretraining for 2000 epochs per layer with GPU acceleration yields a stable embedding $\mathbf{v} = \mathbf{h}_2 \in \mathbb{R}^{20}$. This step effectively distills raw vibrations into a concise set of latent features that emphasize characteristic defect patterns while suppressing noise. An illustrative visualization of the learned sparse patterns is shown in Fig. 2.

C. Profile registration and curvature-driven feature points

Profile contours also require preprocessing, as they are subject to shifts and rotations. Given measured points $\{(x_i, y_i)\}_{i=1}^M$ and a standard template P_r, we solve a rigid registration problem:

$$(R^*, \mathbf{t}^*) = \arg\min_{R \in SO(2), \mathbf{t} \in \mathbb{R}^2} \sum_{i=1}^{M} \|R\mathbf{p}_i + \mathbf{t} - \Pi_{P_r}(\mathbf{p}_i)\|_2^2, \tag{7}$$

where $\mathbf{p}_i = [x_i, y_i]^T$ and Π_{P_r} is the projection operator. To improve robustness, we exploit curvature information. The curvature κ of a smooth parametric curve $\gamma(s) = (x(s), y(s))$ is

$$\kappa(s) = \frac{|x'(s)y''(s) - y'(s)x''(s)|}{\big(x'(s)^2 + y'(s)^2\big)^{3/2}}, \tag{8}$$

which can be approximated by local polynomial fitting. High-curvature points serve as natural landmarks, anchoring registration even in the presence of wear. A validity mask ($y > 140$, $x < 28$) filters irrelevant regions, and interval sampling ensures computational efficiency by reducing redundant points.

D. DTW wear-sequence construction

Once contours are aligned, we quantify localized deviations relative to the template. Let $\mathbf{q}_{1:A}$ denote the experimental contour and $\mathbf{r}_{1:B}$ the reference contour. Dynamic Time Warping (DTW) constructs an accumulated cost matrix:

$$\delta(i, j) = \|\mathbf{q}_i - \mathbf{r}_j\|_2, \tag{9}$$

$$D(1,1) = \delta(1,1), \quad D(i,1) = \delta(i,1) + D(i-1,1), \ i \geq 2, \tag{10}$$

$$D(1,j) = \delta(1,j) + D(1,j-1), \ j \geq 2, \tag{11}$$

$$D(i,j) = \delta(i,j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}. \tag{12}$$

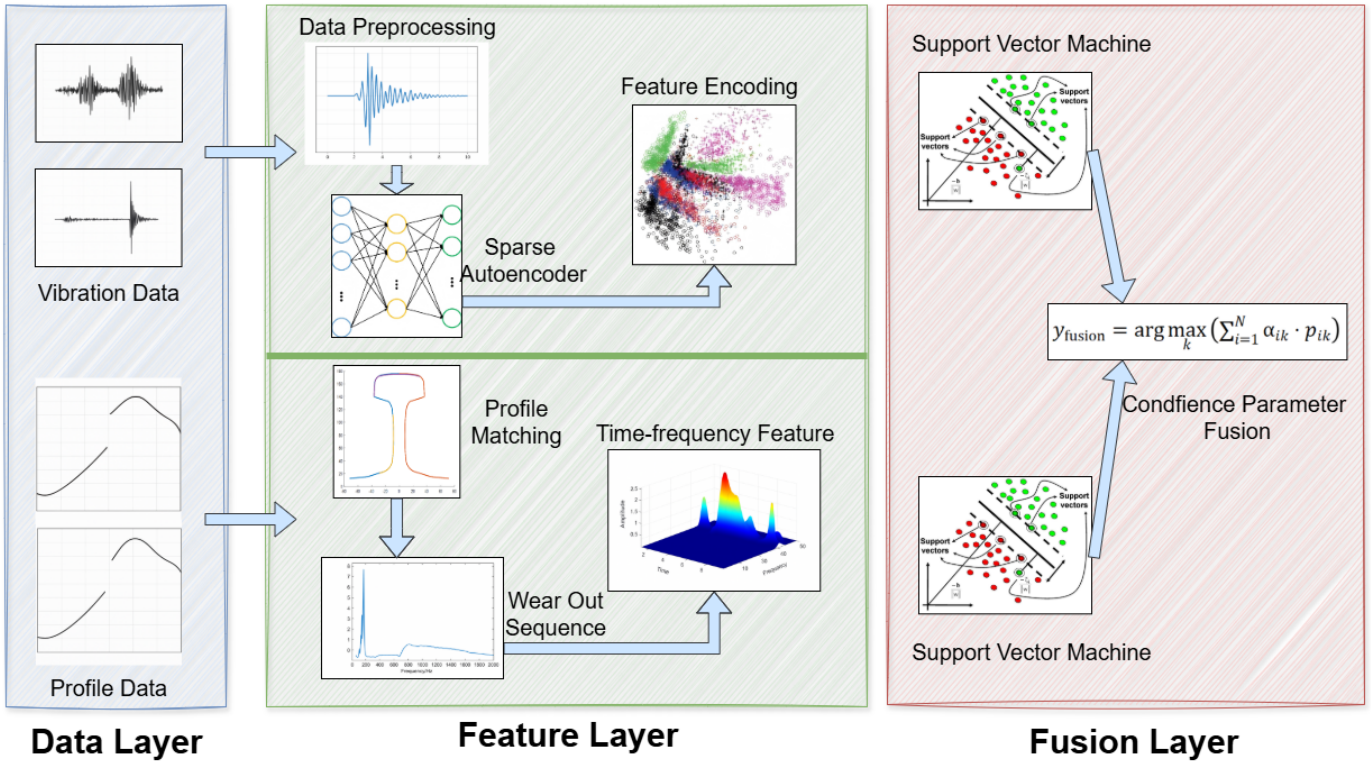

Fig. 1. Overall pipeline: (left) vibration pathway via SAE; (right) profile pathway via registration, curvature features, DTW wear sequence, PCA; (bottom) decision-level class-dependent fusion.

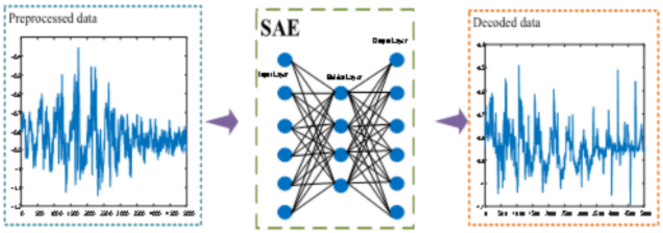

Fig. 2. Sparse stacked autoencoder visualization (kept as in the original).

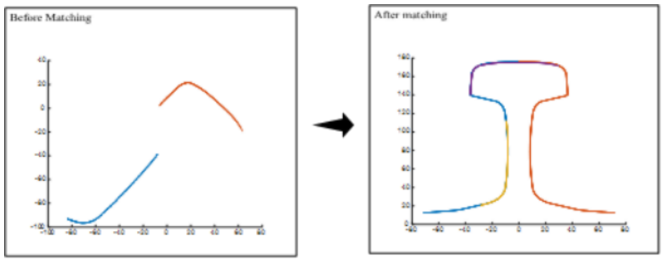

Fig. 3. Profile registration and matching results (kept as in the original).

with $D(1,1) = \delta(1,1)$. By following the path $\mathcal{P}$, we obtain a sequence of pointwise deviations, which we call the *wear sequence* $\mathbf{w} \in \mathbb{R}^{L_w}$. This representation preserves both geometric fidelity and relative alignment, making it particularly suitable for subsequent statistical reduction. Representative registration and matching results on our dataset are shown in Fig. 3.

E. PCA reduction for geometric descriptors

The wear sequences are still high-dimensional, so we employ PCA for compression. Let $W \in \mathbb{R}^{N \times L_w}$ be the centered data matrix. Its covariance is

$$\Sigma = \frac{1}{N-1} W^T W. \tag{13}$$

We then solve the eigenproblem

$$\Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad \lambda_1 \geq \lambda_2 \geq \cdots, \tag{14}$$

and retain the top-K eigenvectors $U_K = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$. Each wear sequence is thus represented as $\mathbf{g} = W U_K \in \mathbb{R}^K$ with K=11 chosen to explain at least 95% of variance.

This step condenses thousands of contour deviations into a compact descriptor that captures the main modes of wear while discarding noise.

F. Per-modality KSVMs and probabilistic outputs

For each modality, we train an RBF-kernel SVM:

$$K(\mathbf{a}, \mathbf{b}) = \exp\Big(-\frac{\|\mathbf{a}-\mathbf{b}\|_2^2}{2\sigma^2}\Big), \quad \sigma > 0. \tag{15}$$

This nonlinear kernel effectively handles the complex decision boundaries required for defect classification. To support decision fusion, we calibrate the raw SVM scores into probabilities using Platt scaling:

$$p_{m,k}(\mathbf{x}) = \sigma_{\log}\big(\alpha_{m,k} f_{m,k}(\mathbf{x}) + \beta_{m,k}\big), \quad \sigma_{\log}(z) = \frac{1}{1+e^{-z}}. \tag{16}$$

Here m indexes the modality (vibration or geometry), and k indexes the defect class. These probabilities represent the classifier's confidence and form the basis of the fusion mechanism.

G. Class-dependent decision fusion

To exploit complementary strengths, we adopt class-dependent fusion. Let $\omega_{m,k} \geq 0$ denote the reliability of modality m for class k, estimated from validation. The fused decision is

$$\hat{y} = \arg\max_k \sum_{m\in\{\text{vib,geo}\}} \omega_{m,k}\, p_{m,k}(\mathbf{x}). \tag{17}$$

In practice, we use $\boldsymbol{\omega}_{\text{geo}} = [0.2, 0.6, 0.7, 0.9]$ and $\boldsymbol{\omega}_{\text{vib}} = [0.6, 0.4, 0.3, 0.1]$, reflecting that geometric descriptors are highly reliable for normal states, whereas vibration cues better capture dynamic signatures such as spalling.

Algorithm 1 Training & Fusion Inference

Input: Paired data $\{(\mathbf{x}^{(n)}, \mathbf{p}^{(n)}, y^{(n)})\}$, PCA target K, KSVM grids C, Σ
Output: Fused classifier $\hat{y}(\cdot)$
1: Normalize $\mathbf{x}^{(n)}$, select $\mathcal{W}_n$, build segments $\mathbf{z}_0^{(n)}$
2: Train SAE (100→20) by minimizing $\mathcal{L}_{\text{SAE}}$; get $\mathbf{v}^{(n)}$
3: Register $\mathbf{p}^{(n)}$ to template; compute wear sequences $\mathbf{w}^{(n)}$
4: Fit PCA on $\{\mathbf{w}^{(n)}\}$; get $\mathbf{g}^{(n)} \in \mathbb{R}^{11}$
5: Train KSVM_{vib} on $\mathbf{v}^{(n)}$; KSVM_{geo} on $\mathbf{g}^{(n)}$
6: Calibrate $p_{\text{vib},k}$ and $p_{\text{geo},k}$ with Platt scaling
7: Tune $\omega_{m,k}$ on validation (grid search)
8: **Inference:** compute $\hat{y} = \arg\max_k \sum_m \omega_{m,k} p_{m,k}(\mathbf{x})$

V. Theoretical and Practical Analysis

A. Why fusion helps: a bias–variance view

The benefit of combining vibration and profile modalities can be formally understood from a classical bias–variance decomposition. Let h_m be the per-modality classifier and $y \in \{1..4\}$. For one-vs-rest decoding, the expected error can be decomposed as

$$\mathbb{E}[\ell(\hat{y}, y)] \approx \sum_m \alpha_m \underbrace{\text{Bias}(h_m)^2}_{\text{systematic}} + \beta_m \underbrace{\text{Var}(h_m)}_{\text{instability}} - \gamma\, \text{Cov}(h_{\text{vib}}, h_{\text{geo}}), \tag{18}$$

where the first term captures systematic misestimation of decision boundaries, the second term reflects random variability due to finite samples, and the last term incorporates cross-modal correlation. Importantly, the negative covariance term (i.e., complementarity) reduces the fused error whenever two classifiers make errors in different regions of the feature space. In practice, axle-box acceleration (ABA) and profile data exhibit distinct sensitivities: vibration is more dynamic but noisy, whereas profile is geometrically precise but less responsive to transient irregularities. Since their error patterns are weakly correlated, weighted fusion is able to simultaneously lower variance and reduce joint confusion, thereby explaining the empirical gains observed in experiments.

B. DTW stability to local misalignment

Another theoretical concern is whether the wear-sequence construction via DTW is sensitive to local misalignment of contours. Let ϕ be a small reparameterization of arclength, which may arise from sampling irregularities, sensor latency, or measurement jitter. For Lipschitz-continuous contours $\mathbf{q}, \mathbf{r}$, one can bound the perturbation in DTW distance as

$$\left|\text{DTW}(\mathbf{q}\circ\phi, \mathbf{r}) - \text{DTW}(\mathbf{q}, \mathbf{r})\right| \;\leq\; \text{Lip}(\mathbf{q}) \cdot \|\phi - \text{id}\|_\infty \cdot |\mathcal{P}|, \tag{19}$$

where $\|\phi - \text{id}\|_\infty$ measures the maximum deviation of ϕ from the identity mapping (i.e., the largest temporal misalignment), $\text{Lip}(\mathbf{q})$ is the Lipschitz constant of the contour $\mathbf{q}$, and $|\mathcal{P}|$ denotes the length of the optimal DTW alignment path (upper bounded by $A + B - 1$ for sequences of length A and B).

This corrected inequality highlights that the variation in DTW cost grows linearly with both the severity of reparameterization and the complexity of the alignment path. In practical terms, it implies that the geometric representation extracted from DTW remains stable under small sampling perturbations. Even if a laser profiler introduces minor shifts or nonuniform spacing, the effect on the final wear-sequence features is bounded and controlled, thereby ensuring the robustness of the profile-based descriptors.

C. Complexity

Finally, we analyze computational complexity to evaluate real-time feasibility. Let L be the selected vibration segment length and M the number of profile points per sample. SAE training requires

$$O(2000 \cdot L \cdot 100 + 2000 \cdot 100 \cdot 20),$$

reflecting two layers with 2000 iterations each. DTW alignment operates with complexity $O(AB)$ where $A, B \sim M$, essentially quadratic in contour length. PCA reduction involves computing the covariance and eigen-decomposition, with cost $O(M^2N + M^3)$. Finally, KSVM training scales between $O(N^2)$ and $O(N^3)$ depending on implementation and kernel approximation. Although some modules are computationally demanding during offline training, all stages are linear or quadratic in practical ranges of N, M, and L. Once trained, the system runs with low per-sample inference cost, making onboard deployment feasible for periodic monitoring and near real-time decision-making.

VI. Experimental Setup

A. Testbed and sensors

To evaluate the proposed approach under controlled yet realistic conditions, we constructed a 100 m experimental rail track in laboratory settings. The rail is designed to include four representative categories of surface states: prefabricated grinding (F1), spalling (F2), abrasion (F3), and intact normal rail (F4). Each defect type was artificially introduced with controlled dimensions and verified by expert inspection to ensure consistency with typical field conditions.

A mobile inspection platform was developed to traverse the rail at a constant, moderate speed. The platform integrates two complementary sensing modules: (i) an axle-box accelerometer (ABA) unit rigidly mounted to capture vertical vibration responses induced by wheel–rail interaction, and (ii) a high-precision laser profile scanner capable of reconstructing the two-dimensional cross-sectional contour of the rail head [12,

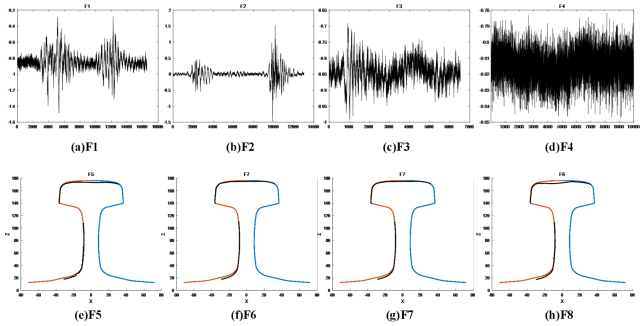

Fig. 4. Representative signals (F1–F4): top row: ABA time-domain waveforms; bottom row: registered rail profiles.

13]. The two sensing channels are hardware-synchronized such that vibration signals and profile measurements are aligned in both space and time [14]. This guarantees that each collected sample corresponds to the same physical segment of the rail, thereby eliminating ambiguities caused by asynchronous acquisition. The laboratory environment also allows control of environmental noise and operating conditions, which facilitates reproducibility of results.

B. Dataset

Based on the above setup, we collected a total of 200 paired samples evenly distributed across the four classes (F1–F4). Each paired sample consists of a time-domain ABA sequence and a corresponding rail profile contour, both aligned to the same spatial location. The collection process strictly enforced spatiotemporal alignment, ensuring that vibration dynamics and geometric deviations are truly paired.

Prior to modeling, all raw signals were pre-processed through several steps: outlier removal to discard corrupted readings, normalization to remove amplitude bias between runs, and ordering to maintain consistent temporal and spatial indices across the dataset. Fig. 4 illustrates representative examples from the dataset: the upper row shows typical vibration waveforms for the four classes, while the lower row depicts the corresponding registered rail profiles. Together, these paired examples highlight both the dynamic and geometric manifestations of the considered short-wavelength irregularities.

C. Implementation details

For the vibration pathway, the stacked autoencoder (SAE) was configured with hidden sizes of $(100, 20)$ neurons, and each layer was pretrained for 2000 iterations to ensure convergence. The learned embeddings provide a compact yet expressive representation of ABA sequences.

For the profile pathway, registered contours were aligned to a standard reference, and wear sequences were constructed using Dynamic Time Warping (DTW) with Euclidean distance as the local cost function. Principal Component Analysis (PCA) was then applied, retaining components that explain at least 95% of the variance, which corresponded to an 11-dimensional geometric descriptor.

Kernel SVMs (KSVMs) with radial basis function (RBF) kernels were trained independently for the vibration and profile modalities. Hyperparameters (C, σ) were selected through grid search on a validation set. Finally, a decision-level fusion strategy was employed, where class-dependent weights were tuned to reflect the relative reliability of each modality. The optimal values were found to be

$$\boldsymbol{\omega}_{\text{geo}} = [0.2, 0.6, 0.7, 0.9], \quad \boldsymbol{\omega}_{\text{vib}} = [0.6, 0.4, 0.3, 0.1],$$

indicating that geometric descriptors contribute more to classes with distinctive shape variations (e.g., F4), while vibration embeddings provide higher discriminative power for dynamic defects such as spalling or abrasion.

VII. Results and Analysis

A. Ablation: vibration-only

With 20D SAE features, KSVM shows limited generalization for F1 and F4 due to overlap with F3. Test accuracy: F1 52.4% (47.6% misclassified as F3), F2 100%, F3 80%, F4 46.4%; metrics in Fig. 5.

These results suggest that vibration features alone, although sensitive to dynamic excitations, do not provide sufficient discriminative power when defects produce partially overlapping frequency or amplitude signatures. In particular, grinding (F1) and abrasion (F3) both introduce high-frequency vibration components that are difficult to separate without additional contextual information. Similarly, the normal rail (F4) sometimes exhibits vibration patterns close to those of abrasion due to environmental noise or minor irregularities that do not qualify as defects but still perturb the ABA signal. This explains the poor recognition rate for F4.

Overall, the vibration-only model excels at spalling (F2), which has a unique and abrupt dynamic signature, but struggles with subtler conditions. This highlights the intrinsic limitation of relying solely on ABA sensing.

B. Ablation: profile-only

After registration→DTW→PCA (11D), KSVM accurately recognizes F4 (100%) and F2 (92.9%), but confuses F1 with F2 (F1 68.8%) and F3 with F4 (F3 77.8%); see Fig. 6.

Compared with vibration-only, the profile-only pathway demonstrates stronger capability in capturing geometric differences, particularly for normal segments (F4), which are sharply separated from defective cases. This confirms that contour information is a highly reliable indicator of whether a rail is defect-free. However, the profile-only model has its own limitations: the confusion between grinding (F1) and spalling (F2) suggests that their wear patterns may share similar cross-sectional characteristics, making it difficult to differentiate based solely on geometry. Likewise, the overlap between abrasion (F3) and normal rails (F4) indicates that slight wear can resemble healthy profiles when inspected only at a geometric level.

Thus, while geometric features capture shape deviations effectively, they are insufficient for capturing the temporal dynamics that distinguish different degradation processes.

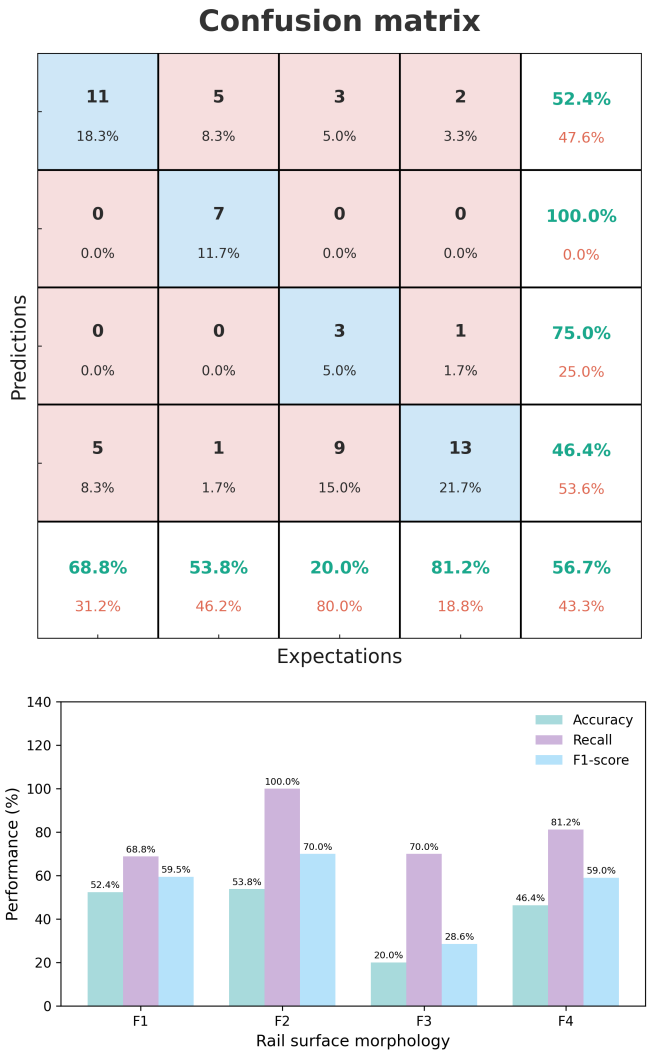

Fig. 5. Vibration-only: (top) confusion matrix; (bottom) precision/recall/F1 across F1–F4.

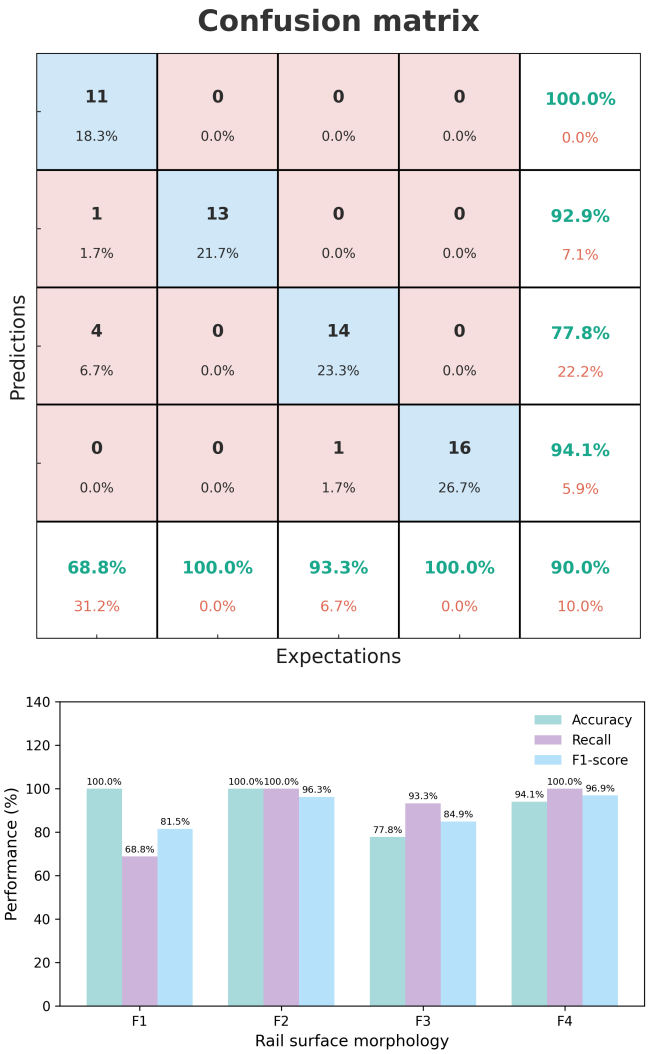

Fig. 6. Profile-only: (top) confusion matrix; (bottom) precision/recall/F1 across F1–F4.

C. Fusion: class-dependent confidence

Fusion substantially reduces F1↔F3 confusions and improves balance: F1 87.5%, F2 100%, F3 93.3%, F4 94.1% (Fig. 7). Table I compares modalities.

The fusion strategy demonstrates the advantage of combining complementary modalities. By weighting vibration more heavily for dynamic defects such as spalling and abrasion, and relying on profile information for distinguishing normal and grinding states, the fused system achieves balanced recognition across all four classes. Importantly, the severe confusions observed in single-modality settings are largely eliminated. The confusion matrix shows that the fusion approach significantly improves the recognition of F1 and F3, which were previously the most difficult to separate.

These improvements validate the hypothesis that ABA and profile data provide orthogonal information. Their errors are weakly correlated, so the fusion model benefits from error compensation, leading to more robust predictions in diverse conditions.

TABLE I Test Accuracy (%) by Modality

Class	Vibration	Profile	Fusion
F4 (Normal)	46.4	**100.0**	94.1
F1 (Grinding)	52.4	68.8	**87.5**
F2 (Spalling)	**100.0**	92.9	**100.0**
F3 (Abrasion)	80.0	77.8	**93.3**

D. Sensitivity and robustness

SAE dimension. Increasing the second-layer size beyond 20 marginally improves training but saturates validation, indicating 20D is near-optimal. This suggests that the learned vibration embeddings already capture the essential dynamic features, and further expansion risks overfitting without clear benefit.

PCA components. Retaining 11 PCs (95%) balances accuracy and complexity; fewer PCs degrade F1/F3 separation. This indicates that geometric variations relevant for distinguishing grinding from abrasion require a relatively rich subspace representation, while further components mostly encode noise.

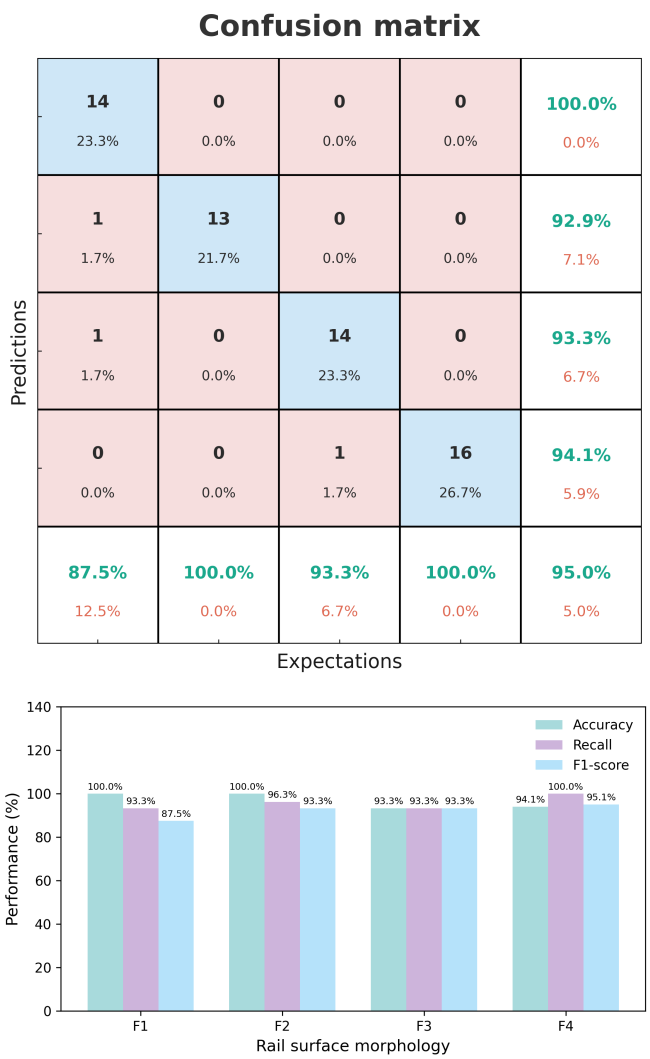

Fig. 7. Fusion: (top) confusion matrix; (bottom) precision/recall/F1 across F1–F4.

Fusion weights. Small perturbations around the selected ω leave accuracy stable, confirming robustness of class-dependent weighting. This stability is critical for real deployment, since it indicates that the fusion model does not rely on overly precise tuning and can tolerate moderate fluctuations in data distribution without significant degradation.

E. Engineering implications

Fusion cuts both false alarms (F4→F3) and misses (F1→F3), enabling more reliable maintenance scheduling and curbing unnecessary grinding. From an engineering standpoint, this lets rail operators avoid costly, unwarranted interventions without compromising safety. A lower false-alarm rate translates directly into savings in inspection time and maintenance budget, while improved detection of subtle defects ensures early action before minor irregularities escalate into severe failures.

Furthermore, the system's robustness enables seamless integration into onboard monitoring platforms with minimal calibration effort, delivering real-time defect identification during normal train operations. This paves the way for intelligent, condition-based maintenance regimes, in place of traditional schedule-based approaches that can either under-maintain or over-maintain infrastructure.

VIII. Discussion

Why F1 and F3 are confusable. Grinding and abrasion produce overlapping excitations at specific speed–wavelength combinations, explaining ABA ambiguity; their cross-sectional wear can also align at shallow depths, explaining profile ambiguity. Fusion integrates dynamics with geometry. In detail, grinding often leaves shallow, periodic marks that generate vibration components resembling those from gradual rail-head abrasion. At certain train speeds, their ABA spectra overlap strongly, making vibration-only classifiers hard to separate. Geometrically, both appear as surface material loss, and their signatures are not easily distinguishable when wear is small. This dual ambiguity explains why single-modality methods fail. By fusing vibration and profile cues—ABA emphasizing transient dynamic excitation and profiles capturing static shape differences—the model breaks the ambiguity.

Deployment. Offline training is modest; online inference is real time. Onboard integration is feasible: continuous ABA streaming with periodic profile scans yields robust alarms at low bandwidth. From an engineering standpoint, ABA can be streamed continuously at low sampling cost, enabling real-time monitoring with minimal compute. Laser profile measurements, though costlier, can run at lower cadence—scheduled at maintenance intervals or triggered by suspicious vibration patterns—balancing cost and accuracy. Crucially, the decision-level fusion requires only lightweight computations at inference, making integration into existing onboard monitoring systems straightforward. Compared with traditional visual inspection or manual ultrasonic testing, this approach substantially reduces human effort and provides continuous, automated coverage, positioning the method as a key component of future intelligent railway maintenance.

Limitations. The lab dataset is controlled; future work should address varied rail types, speeds, environmental conditions, and long-term drift. While a controlled setting ensures consistency, it does not capture field complexity: diverse rail steels, heterogeneous ballast stiffness, weather-driven variability, and rolling-stock noise can affect both vibration and profile signals. Long-term monitoring will also encounter sensor drift, gradual track-geometry changes, and evolving defect characteristics. Addressing these issues will require domain adaptation, transfer learning, and periodic model recalibration. Dataset size is another limitation: although 200 paired samples support a proof of concept, larger, multi-line datasets are needed for statistical generalization. Future studies should also explore deep sequential models or graph-based learning to capture higher-order spatiotemporal correlations in rail degradation. Ultimately, scaling from laboratory demonstration to industrial deployment demands robust validation across diverse rail networks at operational speeds.

IX. Conclusion

This study proposes a multi-modal fusion framework that combines axle-box acceleration (ABA) dynamics with rail geometry to robustly identify short-wavelength irregularities (SWRI). On a 100 m experimental rail with 200 paired samples, the framework employs an SAE+DTW+PCA feature-extraction pipeline together with a class-weighted KSVM for decision fusion, yielding substantial gains. On the test set, it improves accuracy over single-modality baselines by 7.1%–47.7%; the per-class accuracies reach 100% for spalling and 94.1% for normal, while corrugation and abrasion rise to 87.5% and 93.3%, respectively. More importantly, the fusion approach addresses confusions typical of single modalities, significantly reducing misclassification between corrugation and abrasion and narrowing the ambiguous boundary between abrasion and normal.

These improvements stem from the complementary strengths of the two modalities. Vibration sensing is flexible and real-time but susceptible to noise; geometric inspection is precise yet insensitive to transient excitations. By fusing them, the system retains the sensitivity of dynamic responses and the accuracy of geometric measurements, resulting in a more balanced and reliable recognizer. Practically, by lowering false alarms and enabling earlier detection of hazardous defects, the framework can reduce maintenance costs and enhance operational safety. Its lightweight fusion computation makes onboard real-time deployment feasible, laying the groundwork for integration into intelligent inspection vehicles and even next-generation train control systems, and supporting a shift from schedule-based to condition-based maintenance.

That said, scaling to operational networks remains challenging. Larger datasets spanning multiple rail types and operating conditions are needed to verify generalization. Advanced learning methods—such as graph neural networks and self-supervised representation learning—may further improve robustness to unseen defect patterns. Incorporating long-term monitoring data will also allow the system to capture temporal evolution, moving from purely diagnostic outputs toward predictive maintenance.

Overall, this work advances the application of multi-modal learning in infrastructure monitoring and offers a practical solution for railway safety assurance. The successful validation of the fusion framework demonstrates that combining heterogeneous sensing modalities is an effective route to better defect detection, and the approach is readily extensible to other structural health monitoring domains—such as bridges, pavements, and rotating machinery—supporting the development of smarter and safer transportation infrastructure.

Appendix A: Additional Derivations

A.1 SVM primal–dual

One-vs-rest SVM solves

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i \ \ \text{s.t. } y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i,\ \xi_i \geq 0. \tag{20}$$

Dual variables $\boldsymbol{\alpha}$ yield the RBF decision $f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$.

A.2 Curvature via least-squares circle fit

For points $\{(x_i, y_i)\}$ near a head/waist arc, the circle $(x-a)^2 + (y-b)^2 = r^2$ is fit by minimizing

$$\min_{a,b,r} \sum_i \left((x_i-a)^2 + (y_i-b)^2 - r^2\right)^2, \tag{21}$$

whose solution seeds curvature estimates $\kappa = 1/r$.

A.3 Generalization sketch

Under Tsybakov margin with exponent κ, calibrated probabilistic fusion with bounded weights admits

$$\mathcal{R}(\hat{y}) - \mathcal{R}(y^*) \leq \tilde{\mathcal{O}}\Big(\sum_m \tfrac{1}{\sqrt{n_m}}\Big), \tag{22}$$

suggesting that complementary modalities reduce finite-sample error.

References

[1] J. Liu, "Development strategies of railway transportation dispatching modernization," *Railway Transport and Economy*, 2025, 47(2): 1-5, 24.

[2] Y. Zhang, *Study on Vibration Characteristics and Power Spectral Density of Short-wave Irregularities in High-speed Railway Rails*, Dissertation, China Academy of Railway Sciences, Beijing, 2024.

[3] L. j. Shen, C. X. Tian, S. Z. He, et al. "Analysis of short-wave irregularity characteristics of urban rail transit track surface," *Journal of Railway Engineering Society*, 2024, 41(10): 24–30.

[4] X. D. Xu, L. B. Niu, S. C. Sun, et al. "Evaluation method and application of rail corrugation based on axle-box vibration acceleration," *China Railway Science*, 2021, 42(6): 18–26.

[5] L. D. Wang, "The Fast Testing Method of High Speed Railway Rail Corrugation on the Basis of Vibration Response," *China Railway*, 2017(7): 44–49.

[6] K. D. Ahmadi, A. J. Rashidi, A. M. Moghri, et al. "Design and simulation of autonomous military vehicle control system based on machine vision and ensemble movement approach," *Journal of Supercomputing*, 2022, 78: 17309–17347.

[7] H. Ouanan, E. H. Abdelwahed, "Image processing and machine learning applications in mining industry: Mine 4.0," *International Conference on Intelligent Systems and Advanced Computing Sciences*, 2019: 1–5.

[8] A. Esteva, K. Chou, S. Yeung, et al. "Deep learning-enabled medical computer vision," *npj Digital Medicine*, 2021, 4: 5.

[9] T. Wang, B. Chen, Z. Zhang, et al. "Applications of machine vision in agricultural robot navigation: A review," *Computers and Electronics in Agriculture*, 2022, 198: 107085.

[10] Y. Min, B. Xiao, J. Dang, et al. "Real time detection system for rail surface defects based on machine vision," *EURASIP Journal on Image and Video Processing*, 2018, 2018: 41.

[11] C. G. He, K. X. Zhang, R. X. Yu, et al. “Research on the application of machine vision in rail surface disease detection,” *Advanced Materials of High Speed Railway*, 2024, 3(1): 7–13.

[12] H. Wang, *Research and Application of Dynamic Measuring Method of Rail Profile Based on Structured Light Projection*, Dissertation, China Academy of Railway Sciences, Beijing, 2018.

[13] S. N. Wu, *Research on Dynamic Measurement Technology of Rail Wear Based on Rail Vehicle-mounted Line Structured Light Vision*, Dissertation, Zhejiang University, Zhejiang, 2022.

[14] S. B. Zhong, T. Liang, N. Wang, et al. “Application of Rail Full Sectional Profile Detection System in Urban Rail Transit Rail Wear Dynamic Detection,” *Urban Mass Transit*, 2023, 26(10): 143–147.

[15] Z. J. Ye, J. Wang, Y. Wu, “Research on online recognition algorithm of rail profile type based on GA-SVM,” *Instrument Technique and Sensor*, 2024(9): 99–105.

Yiming Zhai is a master student in Control Engineering at Nanjing University of Science and Technology. Her research interests include predictive maintenance and equipment health management.

Huanyu Yang is with the College of Automation, Nanjing University of Science and Technology. His interests include signal processing and multi-modal learning.

Jun Wang is with the College of Automation, Nanjing University of Science and Technology. His interests include structural health monitoring.

Yuntong An received the B.S. degree in automation (railway signal) from Beijing Jiaotong University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Automation and Intelligence.,His research interests include fault diagnosis and condition monitoring in train control systems, and life prediction for railway key equipment.