

Damage detection of steam turbine blades using deep neural network-based acoustic fingerprint recognition

Zhixiang Zhang¹, Wenyu Gao², Yizhuo Gong³, Yunhao Wang⁴, Zun Yin⁵, Xiaomeng Li⁶, Damao Yang⁷, Jin Zhou⁸, Jun Peng⁹

^{1, 2, 3, 4, 8, 9}School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

⁵Xi'an Thermal Power Research Institute Co., Ltd, Xi'an 710054, China

^{6, 7}Huaneng Tongchuan Zhaojin Coal Power Co., Ltd, Tongchuan 727100, China

^{8, 9}Corresponding author

E-mail: ¹1014010236@qq.com, ²gaowenyu037200@163.com, ³lgyz0534@stu.xjtu.edu.cn, ⁴1607147923@qq.com, ⁵yinzun@tpri.com.cn, ⁶ameng12312@126.com, ⁷47773152@qq.com, ⁸jin.zhou@xjtu.edu.cn, ⁹pengjun@xjtu.edu.cn

Received 28 May 2026; accepted 22 August 2026; published online 15 October 2026

DOI <https://doi.org/10.21595/jve.2026.26703>



Copyright © 2026 Zhixiang Zhang, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. This study develops an impact-acoustic framework for classifying predefined surface-damage states on steam-turbine blades, identifying one of three damage regions, and estimating class-dependent severity parameters. Impact signals were recorded for 5 s at 22,050 Hz under controlled impact and microphone angles. After event screening, each signal was converted into a 13-coefficient W-MFCC map that combines continuous-wavelet time-frequency localization, Mel-scale subband mapping, and cepstral compression. A hybrid ResNet-50-Transformer-PANet-CMTL network was designed to extract hierarchical local features, model long-range dependencies, fuse multi-scale information, and produce conditional multi-task outputs. The dataset contained 1,200 training samples and 300 validation samples covering crack, crack-erosion, and crack-abrasion conditions in Areas 1-3. On the reported validation split, W-MFCC correctly classified 296 of 300 samples (98.67 %), compared with 89.33 % for MFCC and 87.33 % for wavelet input, and its loss curves showed more stable convergence. The full multi-scale configuration also exhibited smoother training behavior than the ResNet-50 + Transformer baseline. These findings demonstrate the feasibility of combining W-MFCC representation learning with multi-scale conditional prediction for controlled turbine-blade condition assessment.

Keywords: acoustic signature, steam-turbine blade, damage classification, W-MFCC, multi-task learning.

1. Introduction

Steam turbines are central energy-conversion systems in thermal power generation and are critical to reliable power supply [1, 2]. Their blades operate under centrifugal loading, thermal stress, aerodynamic vibration, wet-steam erosion, and repeated droplet impact. Long-term exposure can produce cracks, abrasion, erosion, and local material loss. Reliable assessment of these conditions is therefore important for maintenance planning and safe operation [3].

Common nondestructive inspection methods for turbine blades include ultrasonic testing, eddy-current testing, magnetic-particle testing, and visual inspection [1, 2, 4, 5]. Ultrasonic methods can detect small flaws, but complex blade curvature, surface roughness, coupling conditions, and nonuniform propagation paths may create blind zones or complicate quantification [3, 6]. Eddy-current methods are also sensitive to probe lift-off, surface geometry, and electromagnetic-field distortion, which can reduce reliability on complex curved surfaces [7-10]. These limitations motivate complementary inspection approaches that can be implemented with compact sensing hardware.

Mechanical impact and vibration generate acoustic signatures containing time-, frequency-, and spatial-domain information. Changes in geometry, boundary conditions, and surface condition can alter these signatures, making them potentially useful for condition classification [11-14].

Related studies have applied acoustic representations to target recognition, fault diagnosis, and near-field acoustic analysis [11-22].

Conventional time- and frequency-domain descriptors may not fully represent transient, nonlinear, and high-dimensional acoustic responses. Compact robotic inspection platforms can carry sensing hardware to otherwise difficult-to-access surfaces [23], increasing the need for informative representations that can be processed efficiently. Deep-learning methods provide a data-driven means of learning discriminative features from such representations [24-26], and recent studies have demonstrated their usefulness for acoustic fault diagnosis and condition monitoring [5, 27-30].

Residual networks mitigate vanishing gradients and degradation through identity-based skip connections, enabling stable optimization of deep feature extractors [31]. Nevertheless, local convolution alone may not capture long-range dependencies in time–frequency maps. Multi-head self-attention can complement convolution by modeling interactions among distant feature regions.

Damage-related patterns may also occur at different spatial and frequency scales. Shallow features retain local detail, whereas deeper features contain stronger semantic information. A single-scale representation can therefore be insufficient for classifying cracks, erosion, abrasion, and mixed conditions.

Transformer modules provide global context through attention, while feature-pyramid architectures combine information across scales. Their roles are complementary rather than interchangeable: attention models long-range dependencies, and multi-scale fusion restores detail that can be weakened by downsampling.

The Feature Pyramid Network uses a top-down pathway to enrich lower-level features [32]. PANet adds a bottom-up aggregation path that shortens information flow and strengthens multi-scale fusion [33-36]. These properties are relevant to acoustic damage representations, where both localized transients and broader spectral patterns may contribute to the predictions.

The novelty of this study lies in the task-specific integration of complementary representation and network components. First, the proposed W-MFCC representation aligns wavelet-localized impact transients with Mel-spaced subband aggregation and cepstral compression, thereby retaining local time–frequency detail while reducing spectral redundancy. Second, multi-head self-attention is applied to the semantically strongest C_5 feature map, while the shallower C_1 – C_4 maps are preserved for multi-scale fusion. Third, PANet combines high-level semantic information with lower-level detail through bidirectional feature aggregation. Finally, the conditional multi-task head links damage-type classification to class-specific severity regression while also identifying the damage region. Unlike a single-representation or single-scale pipeline, the resulting framework jointly models damage type, region, and severity within one controlled impact-acoustic assessment task.

The experimental scope is limited to the predefined crack, crack–erosion, and crack–abrasion conditions represented in the training and validation datasets, together with the associated severity descriptors and Area 1–Area 3 region labels. Accordingly, the reported findings should be interpreted as evidence of feasibility under the controlled acquisition protocol rather than as a general assessment of all turbine-blade damage modes or field operating conditions.

2. Theoretical background

2.1. Acoustic signature theory

An acoustic signature is a signal generated by impact, vibration, or another mechanical excitation. Its waveform depends on material properties, geometry, surface condition, boundary conditions, and the excitation-measurement configuration. Damage can modify the propagation path and modal response, thereby changing time-, frequency-, and energy-related features.

In practice, damage modes may coexist and their acoustic effects may be coupled. A complete

closed-form mapping from each damage geometry to a unique set of acoustic parameters is therefore difficult to establish for complex blades. A data-driven model can instead learn statistical associations between labeled damage conditions and measured acoustic features.

Accordingly, the present framework uses a neural network to learn discriminative acoustic features and jointly perform damage-condition classification, region identification, and severity-related regression.

The microphone was positioned in the near-field region relative to the impact location. In this region, the measured response is sensitive to source geometry, excitation position, and local boundary conditions. The study therefore uses a controlled acquisition geometry to reduce avoidable variability.

2.2. Wavelet-MFCC multi-resolution feature fusion

Deep neural networks can learn nonlinear mappings, but their performance depends strongly on the input representation. For impact-acoustic signals, preprocessing should preserve short transients while reducing redundant information and arranging the frequency content in a form suitable for learning.

The measured signals are nonstationary and contain abrupt impact responses, broad frequency content, and local variations caused by geometry and boundary conditions. Continuous wavelet analysis provides time-scale localization, whereas Mel-frequency cepstral coefficients offer a compact representation of subband energy. The proposed W-MFCC framework combines these complementary properties.

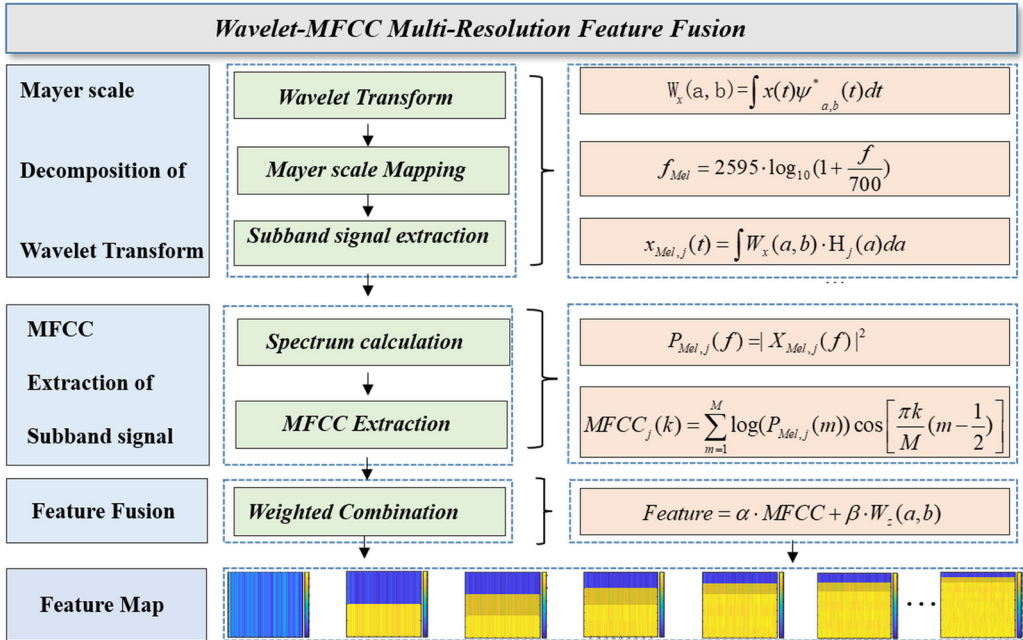


Fig. 1. Schematic of the W-MFCC feature-extraction framework

Fig. 1 provides an overview of the complete W-MFCC feature-extraction pipeline, which consists of four main stages: wavelet analysis and scale-to-frequency mapping, Mel-subband energy extraction, MFCC computation, and feature fusion. The blocks on the left represent these successive processing stages, while the band at the top indicates the overall workflow. The color maps on the right show the normalized wavelet-energy distribution, the Mel-subband log-energy representation, and the final fused W-MFCC features. In these maps, warmer colors correspond

to larger normalized coefficients, and variations in the color distribution reflect differences in energy concentration across time and frequency. These maps are provided solely for schematic illustration and do not represent additional experimental results.

2.2.1. Wavelet decomposition and Mel-scale mapping

The continuous wavelet transform of the signal $x(t)$ is evaluated at scale a and translation b . Here, $a > 0$ is a dimensionless scale parameter that dilates or compresses the mother wavelet, and b is the temporal translation parameter. In a sampled signal, b is evaluated at discrete time positions determined by the sampling interval. The selected scales depend on the sampling frequency, the center frequency of the mother wavelet, and the frequency band of interest. The symbol $*$ denotes complex conjugation:

$$W_x(a, b) = \int x(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t - b}{a} \right) dt. \quad (1)$$

The pseudo-frequency associated with scale a is given by Eq. (2). In this expression, $f(a)$ is the pseudo-frequency represented by scale a , f_ψ is the normalized center frequency of the mother wavelet, f_s is the sampling frequency (22,050 Hz in this study), and a is the wavelet scale. Larger values of a correspond to lower pseudo-frequencies, whereas smaller values correspond to higher pseudo-frequencies:

$$f(a) = \frac{f_\psi f_s}{a}. \quad (2)$$

The frequency axis is then mapped to the Mel scale. The physical frequency f is converted to the Mel coordinate $m(f)$ using Eq. (3):

$$m(f) = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right). \quad (3)$$

Uniform samples on the Mel axis are mapped back to physical frequency using the inverse relation in Eq. (4):

$$f(m) = 700(10^{m/2595} - 1). \quad (4)$$

This procedure aligns wavelet-domain frequency bands with Mel-spaced subbands before energy aggregation.

2.2.2. Subband energy extraction

After Mel-scale resampling, M triangular filters are applied to the wavelet energy spectrum. The energy of the m -th subband at frame position b is defined by Eq. (5):

$$E_m(b) = \sum_k H_m(f_k) \cdot |W_x(f_k, b)|^2. \quad (5)$$

The operation maps multi-scale wavelet energy into a compact perceptual-frequency space while retaining temporal variation.

2.2.3. MFCC calculation

The logarithm of each subband energy is calculated to compress the dynamic range and

improve numerical stability, as shown in Eq. (6):

$$\tilde{E}_m = \log(E_m + \varepsilon). \quad (6)$$

The small constant ε prevents the logarithm of zero. A type-II discrete cosine transform is then applied, and the first N coefficients are retained as MFCC features according to Eq. (7):

$$c_n = \sum_{m=1}^M \tilde{E}_m \cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right], \quad n = 0, \dots, N_c - 1. \quad (7)$$

This step converts the log-energy spectrum into a lower-dimensional cepstral representation.

2.2.4. Multi-feature fusion

Because MFCCs provide compact spectral information but may smooth short transients, additional time-frequency features are extracted from the wavelet domain to form vector g and are fused with the MFCC vector c :

$$z = \alpha \cdot c \oplus (1 - \alpha) \cdot g, \quad 0 \leq \alpha \leq 1. \quad (8)$$

In Eq. (8), $\oplus$ denotes concatenation or weighted combination, and α is the fusion coefficient. The same definition is used throughout the manuscript.

The resulting W-MFCC representation combines wavelet-domain transient information with Mel-frequency cepstral compression.

2.3. ResNet-50 network

The W-MFCC maps are input to a neural network that performs damage-condition classification, region identification, and severity-related regression.

ResNet-50 is used as the convolutional backbone because its residual connections support stable training of a deep feature extractor.

As shown in Fig. 2(a), the network contains an initial convolution and max-pooling stage followed by four residual stages. The standard bottleneck block in Fig. 2(b) contains three sequential convolutions: a 1×1 convolution for channel reduction, a 3×3 convolution for spatial feature extraction, and a 1×1 convolution for channel expansion. The skip connection is a parallel identity or projection path and is not one of the three convolutional layers. This bottleneck design controls parameter growth while enabling a deep network.

The residual mapping is expressed by Eq. (9):

$$H(x) = F(x) + x. \quad (9)$$

The bottleneck residual function shown in Fig. 2(b) is expressed by Eq. (10):

$$F(x) = W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 \cdot x)). \quad (10)$$

The first 1×1 convolution reduces the channel dimension, the 3×3 convolution extracts spatial features, and the final 1×1 convolution restores the output dimension.

2.4. Transformer attention mechanism

The highest-level C_5 feature map is passed to a multi-head self-attention module after channel adaptation and positional encoding. The attention module models relationships among distant

regions of the time-frequency representation, complementing the local receptive fields of the convolutional backbone:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (11)$$

In Eq. (11), Q , K , and V are the query, key, and value matrices. The scaled dot product QK^T determines pairwise similarity, the softmax operation produces normalized attention weights, and the weighted sum of V forms the attention output:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (12)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (13)$$

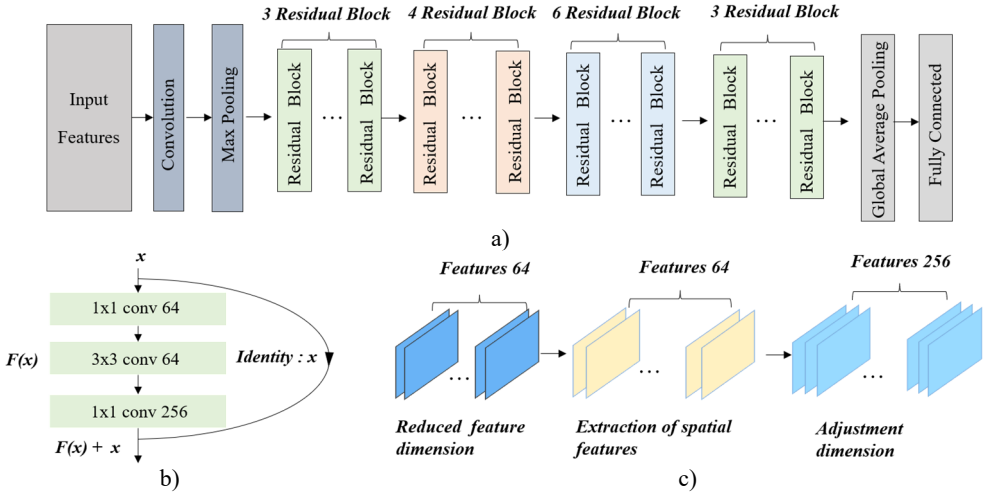


Fig. 2. ResNet-50 architecture: a) overall network with the initial convolution, max pooling, four residual stages, global average pooling, and the output layer, b) bottleneck residual block with 1×1 , 3×3 , and 1×1 convolutions and a parallel skip connection, c) channel reduction, spatial feature extraction, and channel expansion within the bottleneck block

Eqs. (12-13) define the parallel attention heads and their concatenation. Each head uses independent projection matrices and can emphasize a different feature subspace. The current implementation uses eight heads.

3. Proposed method

The proposed architecture combines ResNet-50, multi-head self-attention, PANet, and a conditional multi-task output head. ResNet-50 extracts local and hierarchical features, the attention module models global dependencies, PANet fuses multi-scale information, and the output head predicts damage type, damage region (Area 1-Area 3), and severity-related parameters. Because the region labels are categorical, the corresponding task is defined as region identification rather than coordinate-level localization.

Fig. 3 presents the complete workflow, from acoustic-signal acquisition and preprocessing to feature generation, dataset construction, network training, and multi-task prediction.

Impact-acoustic signals were recorded for 5 s at a sampling rate of 22,050 Hz. The acquisition conditions are summarized in Table 1. An energy-threshold procedure and a sliding window were used to retain valid impact events and suppress low-energy background segments.

Each retained signal was converted to a W-MFCC representation containing 13 cepstral

coefficients. The resulting two-dimensional map was used as the network input.

The feature maps were annotated with damage region, damage type, and the available geometric descriptors. Region labels comprised Area 1-Area 3, whereas the damage classes included crack, erosion, abrasion, and mixed conditions. All annotations were stored in JSON format.

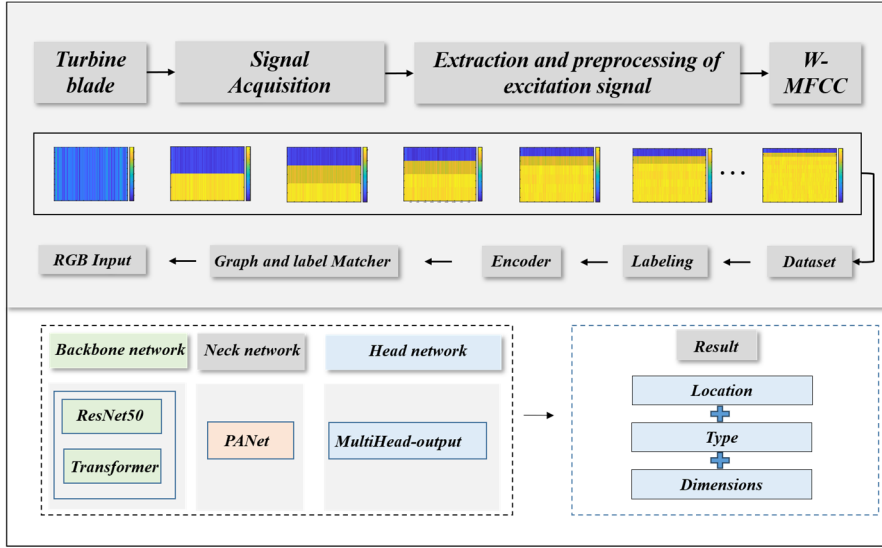


Fig. 3. Overall experimental workflow

Table 1. Data acquisition settings.

Region	Point	Level	Impact angle	Microphone angle
Area 1	Points 1-4	Levels 1-3	0°, 45°, 90°	0°, 45°, 90°
Area 2	Points 1-4	Levels 1-3	0°, 45°, 90°	0°, 45°, 90°
Area 3	Points 1-4	Levels 1-3	0°, 45°, 90°	0°, 45°, 90°

During dataset construction, the JSON annotations were parsed and associated with their corresponding feature maps. The resulting image-label pairs were then encoded for multi-task learning.

The customized MultiTaskDataGenerator assembled the encoded samples into the input tensors and task-specific target tensors required for model training and validation.

3.1. Model construction and selection

The modules were connected so that the convolutional backbone, attention block, multi-scale neck, and conditional output head could exchange features with compatible channel dimensions.

3.1.1. Adaptation of ResNet-50

ResNet-50 provides feature maps C_1 - C_5 . A 1×1 adapter aligns the channel dimensions required by the attention and PANet modules. Only C_5 is passed through self-attention because it has the strongest semantic representation and the smallest spatial size, which reduces attention cost. The shallower C_1 - C_4 features are retained for multi-scale fusion. Eqs. (14-17) describe the dimension adaptation:

$$F_{ResNet} = ResNet50(x), \quad x \in R^{B \times C \times H \times W}, \quad (14)$$

$$F_{ResNet_flattened} = Flatten(F_{ResNet}), \quad (15)$$

$$F_{ResNet_adapted} = Linear(F_{ResNet_flattened}), \quad (16)$$

$$F'_{C_5} = Transformer(F_{ResNet_adapted}). \quad (17)$$

After adaptation, learnable positional encoding is added and the C_5 sequence is processed by eight attention heads using Eqs. (11-13).

3.1.2. PANet neck

The attention-enhanced C_5 feature is resized and integrated with the convolutional feature pyramid:

$$F_{PANet}^{(5)} = Upsample(F'_{C_5}, size = (H_l, W_l)). \quad (18)$$

Eq. (19) describes the top-down pathway. Here, the index l denotes the feature level, the lateral 1×1 convolution aligns channels, and the resized higher-level feature enriches the semantic content of the lower-level feature:

$$F_l^{td} = Conv(F_l + Upsample(F_{l+1}^{td})). \quad (19)$$

Eqs. (20-21) describe the bottom-up aggregation pathway and the final fused features:

$$F_l^{bu} = Conv(F_l^{td} + Downsample(F_{l-1}^{bu})), \quad (20)$$

$$F_{fusion} = Concat(F_2^{bu}, F_3^{bu}, F_4^{bu}, F_5^{bu}). \quad (21)$$

The bidirectional paths combine higher-level semantics with lower-level detail.

3.1.3. Conditional multi-task output head

The Conditional Cascaded Multi-Task Learning (CMTL) head first predicts the damage class and then activates the corresponding regression branch. This sequential design represents the dependency between damage type and its severity-related parameters.

Let p denote the predicted class-probability vector and r_k the output of the regression branch for class k . The classification loss is defined in Eq. (22):

$$l_{cls} = - \sum_{k=1}^K y_{cls}^{(k)} \cdot \log(\hat{y}_{cls}^{(k)}). \quad (22)$$

The regression loss is evaluated for the branch associated with the target class, and the weighted overall loss is defined in Eq. (23):

$$l_{total} = \lambda_1 \cdot l_{cls} + \lambda_2 \cdot l_{reg}. \quad (23)$$

Fig. 4 summarizes the proposed architecture. ResNet-50 extracts multi-level features; the C_5 feature is enhanced by multi-head self-attention; PANet performs top-down and bottom-up fusion; and the conditional multi-task head produces damage-type, damage-region, and severity-related outputs.

3.2. Experiments and results

The data were partitioned into the training and validation sets summarized in Tables 2 and 3,

respectively.

Figs. 5-7 present the comparative results for the three signal representations and the optimization behavior of the baseline and proposed network configurations.

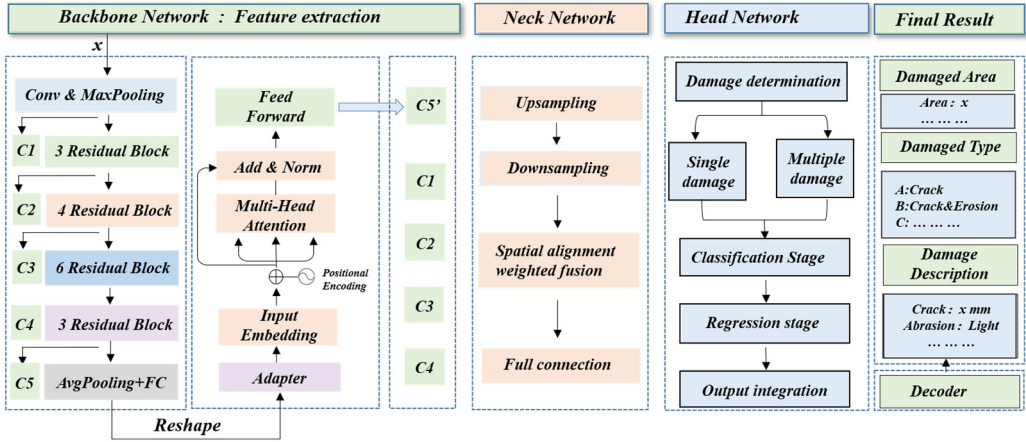


Fig. 4. Architecture of the proposed hybrid network

3.2.1. Data sources and composition

Within each 5 cm×5 cm region, impacts were applied at four points on a 4 cm×4 cm grid. Three impact levels were produced using ball-drop heights of 30 cm, 40 cm, and 50 cm. The training and validation compositions are listed in Tables 2 and 3.

Table 2. Training dataset composition

Region	Damage type	Dimensions (mm)	Samples
Area 1	Crack	Crack: 0.3	400
Area 2	Crack + erosion	Crack: 0.5 Erosion: 1×2×2	400
Area 3	Crack + abrasion	Crack: 0.7 Abrasion: 5×5×0.2	400

Table 3. Validation dataset composition

Region	Damage type	Dimensions (mm)	Samples
Area 1	Crack	Crack: 4.6	100
Area 2	Crack + erosion	Crack: 0.8 Erosion: 2×5×2	100
Area 3	Crack + abrasion	Crack: 3 Abrasion: 8×5×0.3	100

3.2.2. Network and signal-representation results

Fig. 5 compares representative damage conditions and the corresponding MFCC, wavelet, and W-MFCC representations. Figs. 5(a)-5(c) show the damage examples; Figs. 5(d)-5(f), 5(g)-5(i), and 5(j)-5(l) show the three signal representations. The 0.3 mm crack in Fig. 5(a) is included as a representative training example.

Fig. 6 compares the training behavior of the ResNet-50 + Transformer baseline with that of the full ResNet-50 + Transformer + PANet + CMTL architecture. The baseline exhibits pronounced fluctuations and late spikes in the total and task-specific losses, whereas the full architecture converges more smoothly toward lower loss values. Training accuracies for both damage-type classification and region identification also rise to a more stable plateau in the full

model. At each epoch, the two accuracies are calculated as the proportions of correctly classified samples for their respective damage-type and Area 1-Area 3 labels.

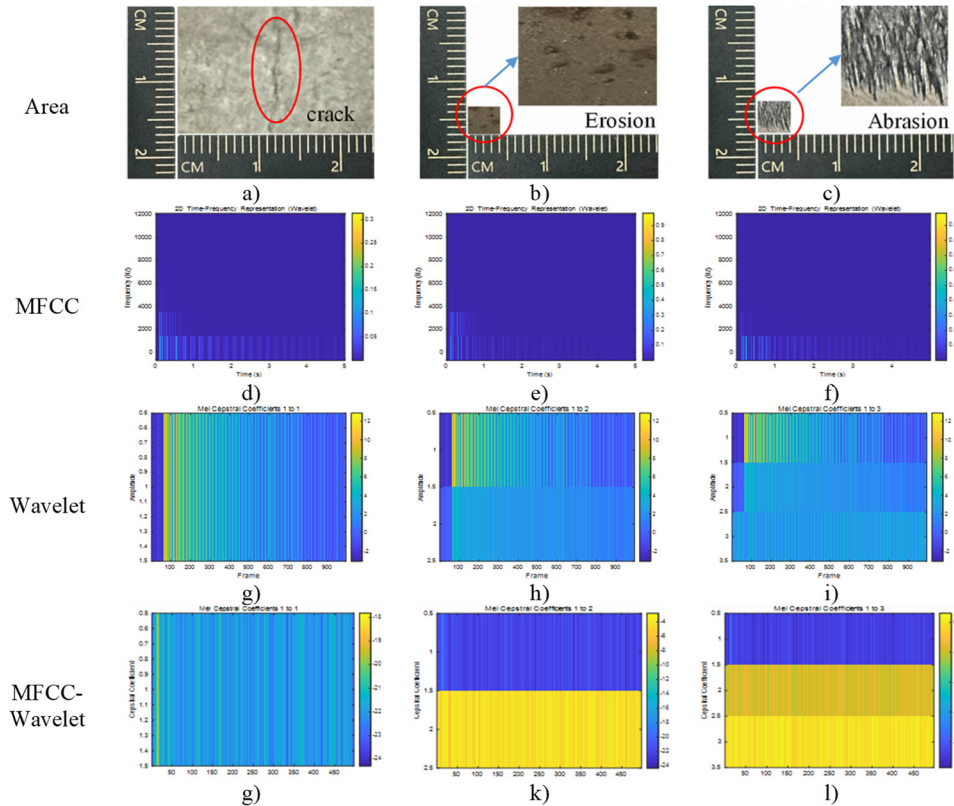


Fig. 5. Representative damage conditions and signal representations: a)-c) crack, erosion, and abrasion examples, d)-f) MFCC representations, g)-i) wavelet representations, j)-l) W-MFCC representations. Photographs in panels a)-c) were taken by the authors during the experimental study at International Center for Applied Mechanics, Xi'an, China, in March 2026

3.2.3. Validation setup and interpretation

Table 2 summarizes the training dataset, whereas Table 3 presents the corresponding validation dataset. Based on these datasets, Fig. 7 compares the confusion matrices, loss curves, and accuracy curves obtained using MFCC, wavelet, and W-MFCC representations as network inputs. Table 4 summarizes damage-classification performance for the three signal representations. Among them, W-MFCC exhibits the most stable decrease in loss and achieves the highest validation accuracy (98.67 %, 296/300). MFCC and wavelet inputs show larger optimization fluctuations, indicating that either representation alone may not fully capture the complex characteristics of impact-acoustic signals generated by damaged turbine blades.

Table 4. Comparison of damage-classification performance using different signal representations

Representation	Correct samples	Total samples	Accuracy
MFCC	268	300	89.33 %
Wavelet	262	300	87.33 %
W-MFCC	296	300	98.67 %

The improved performance of W-MFCC can be attributed to the complementary information provided by wavelet analysis and Mel-frequency cepstral processing. Wavelet analysis offers

strong time-frequency localization and is therefore effective in preserving abrupt, short-duration, and spatially localized variations associated with impact responses and surface damage. However, wavelet coefficients may contain redundant information and can be sensitive to local fluctuations. In contrast, Mel-spaced filtering reorganizes the frequency spectrum into perceptually and structurally meaningful subbands, while logarithmic compression and discrete cosine transformation produce a compact cepstral representation of the overall spectral-energy distribution. MFCC features are therefore efficient for describing global spectral patterns, although some transient details may be weakened during compression.

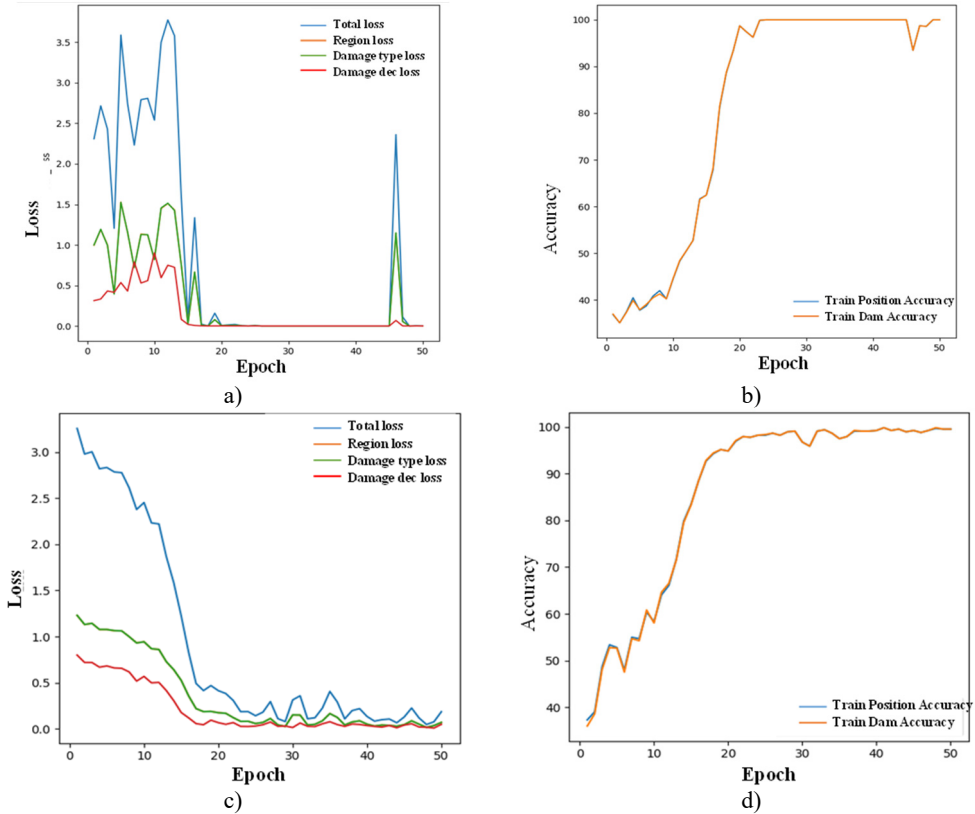


Fig. 6. Training curves of the compared network configurations: a) loss curves and b) training accuracy for the ResNet-50 + Transformer baseline; c) loss curves and d) training accuracy for the full ResNet-50 + Transformer + PANet + CMTL model

By combining these two representations, W-MFCC retains localized transient information while reducing spectral redundancy and improving the organization of the input feature space. This fusion provides the neural network with a more structured and discriminative representation than either MFCC or wavelet features alone. As a result, the model can more effectively learn both local damage-related variations and broader spectral characteristics, leading to smoother convergence, lower optimization instability, and improved classification performance in the reported experiments.

These findings indicate that feature representation plays an important role in acoustic-based turbine-blade condition assessment. The results support the use of W-MFCC as a suitable input representation for the proposed multi-task network under the investigated experimental conditions. Nevertheless, the comparison is based on the available training and validation split, and repeated independent experiments and statistical uncertainty analyses were not conducted. Therefore, further evaluation using additional blades, broader defect dimensions, different

operating and noise conditions, and repeated trials is required to confirm the generalization capability and robustness of the W-MFCC representation.

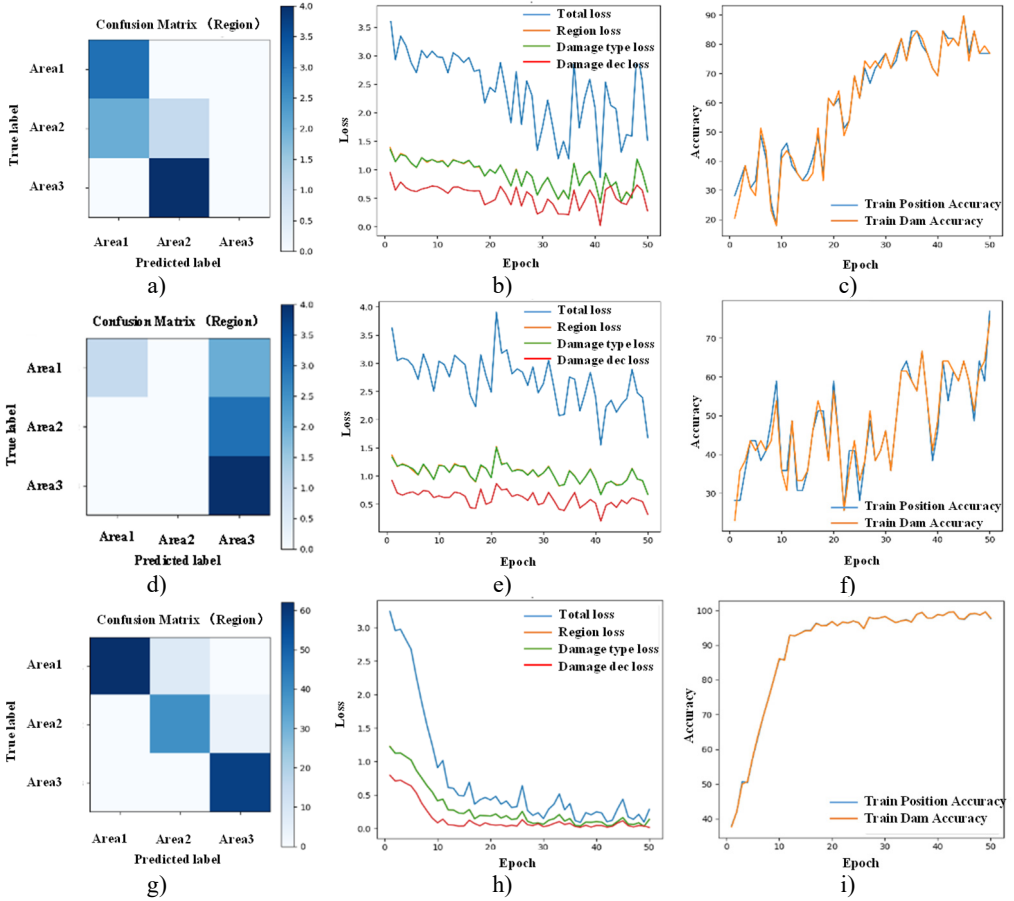


Fig. 7. Results obtained with different signal representations:
 a)-c) MFCC input; d)-f) wavelet input, g)-i) W-MFCC input

4. Discussion

The comparative curves indicate that the W-MFCC representation provides a more stable optimization process than MFCC or wavelet input alone. This behavior may be attributed to the complementary information retained by the two components: wavelet analysis preserves localized transient variations, whereas Mel-frequency cepstral compression provides a compact description of the spectral-energy distribution.

The network components also serve complementary functions. ResNet-50 extracts hierarchical local features, self-attention models long-range dependencies at the C_5 level, PANet integrates semantic and detailed information across scales, and the conditional output head couples damage-type classification with class-specific regression. The methodological contribution therefore lies in adapting and integrating these components within a unified acoustic multi-task framework.

However, the study has several limitations. First, the dataset covers only three predefined regions and a limited range of artificial defect dimensions, so it does not provide a comprehensive assessment across defect scales. Second, blade-independent testing, controlled-noise experiments, repeated training runs, and statistical uncertainty analyses were not conducted. Third, although the validation set contains defect dimensions different from those in the training set, both sets were

acquired under the same controlled protocol. Future work should therefore expand the dataset to include additional blades, broader defect types and dimensions, variable impact and microphone configurations, and representative operating noise. Independent blade-level tests and repeated-run statistical analyses will be required to quantify generalization and robustness.

5. Conclusions

This study developed a W-MFCC representation and a hybrid ResNet-50-Transformer-PANet-CMTL network for classifying predefined steam-turbine blade damage conditions, identifying one of three damage regions, and estimating class-dependent severity parameters. Three main conclusions can be drawn. First, combining wavelet-localized transient information with Mel-frequency cepstral compression produced a more discriminative input representation: W-MFCC correctly classified 296 of 300 validation samples (98.67 %), compared with 89.33 % for MFCC and 87.33 % for wavelet input. Second, the full multi-scale architecture showed smoother loss reduction and more stable training behavior than the ResNet-50 + Transformer baseline, supporting the complementary roles of global attention and bidirectional feature fusion. Third, the conditional output design provides a unified route for damage-type classification, Area 1-Area 3 region identification, and class-specific severity regression under the investigated acquisition conditions. These results establish the feasibility of the proposed framework for controlled impact-acoustic turbine-blade assessment. Nevertheless, the conclusions are limited to the available train-validation split and predefined artificial defects. Future work will extend the study to independent blades, wider defect scales, controlled and operational noise, repeated training trials, and statistical uncertainty analysis before practical field deployment.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (grant number: 52175248).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Jun Peng: project administration and funding acquisition. Jin Zhou: supervision. Zhixiang Zhang: conceptualization, data curation, methodology, software. Wenyu Gao: resources, validation, data curation, formal analysis. Yizhuo Gong: validation, writing-review and editing. Yunhao Wang: writing-review and editing. Zun Yin: writing-review and editing. Xiaomeng Li: writing-review and editing. Damao Yang: writing-review and editing.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- [1] T. Stubbs, "The role of NDE in the life management of steam turbine rotors," *Insight – Non-destructive Testing and Condition Monitoring*, Vol. 46, No. 9, pp. 529–532, Sep. 2004, <https://doi.org/10.1784/insi.46.9.529.40841>
- [2] J. Veselá, P. Mareš, P. Zahrádka, and J. Patera, "Evaluation of steam turbine blades surface cracks detectability by nondestructive methods," *Journal of Nuclear Engineering and Radiation Science*, Vol. 7, No. 2, Apr. 2021, <https://doi.org/10.1115/1.4048479>

- [3] H. Mao et al., "Nonlinear ultrasonic characterization of carburized case depth," *NDT and E International*, Vol. 112, p. 102244, Jun. 2020, <https://doi.org/10.1016/j.ndteint.2020.102244>
- [4] Q. Miao et al., "Creep feed grinding induced gradient microstructures in the superficial layer of turbine blade root of single crystal nickel-based superalloy," *International Journal of Extreme Manufacturing*, Vol. 3, No. 4, p. 045102, Oct. 2021, <https://doi.org/10.1088/2631-7990/ac1e05>
- [5] H. Zhao, Y. Gao, and W. Deng, "Defect detection using Shuffle Net-CA-SSD lightweight network for turbine blades in IoT," *IEEE Internet of Things Journal*, Vol. 11, No. 20, pp. 32804–32812, Oct. 2024, <https://doi.org/10.1109/jiot.2024.3409823>
- [6] F. Mevissen and M. Meo, "A nonlinear ultrasonic modulation method for crack detection in turbine blades," *Aerospace*, Vol. 7, No. 6, p. 72, Jun. 2020, <https://doi.org/10.3390/aerospace7060072>
- [7] A. N. Abdalla, M. A. Faraj, F. Samsuri, D. Rifai, K. Ali, and Y. Al-Douri, "Challenges in improving the performance of eddy current testing: review," *Measurement and Control*, Vol. 52, No. 1-2, pp. 46–64, Jan. 2018, <https://doi.org/10.1177/0020294018801382>
- [8] Y. He, F. Luo, M. Pan, X. Hu, J. Gao, and B. Liu, "Defect classification based on rectangular pulsed eddy current sensor in different directions," *Sensors and Actuators A: Physical*, Vol. 157, No. 1, pp. 26–31, Nov. 2009, <https://doi.org/10.1016/j.sna.2009.11.012>
- [9] J. Wu, D. Zhou, and J. Wang, "Surface crack detection for carbon fiber reinforced plastic materials using pulsed eddy current based on rectangular differential probe," *Journal of Sensors*, Vol. 2014, pp. 1–8, Oct. 2014, <https://doi.org/10.1155/2014/727269>
- [10] H. Zhang, Z. Cheng, H. Wang, and F. Xie, "Research on weld defect detection based on eddy current thermal imaging," in *International Conference on Intelligent Equipment and Special Robots (ICIESR 2021)*, p. 67, 2021, <https://doi.org/10.1117/12.2625295>
- [11] Z. Chen, J. Tang, H. Qiu, and M. Chen, "MGFGNet: an automatic underwater acoustic target recognition method based on the multi-gradient flow global feature enhancement network," *Frontiers in Marine Science*, Vol. 10, Dec. 2023, <https://doi.org/10.3389/fmars.2023.1306229>
- [12] Z. Chen, G. Xie, M. Chen, and H. Qiu, "Model for underwater acoustic target recognition with attention mechanism based on residual concatenate," *Journal of Marine Science and Engineering*, Vol. 12, No. 1, p. 24, Dec. 2023, <https://doi.org/10.3390/jmse12010024>
- [13] J. Jiang, T. Shi, M. Huang, and Z. Xiao, "Multi-scale spectral feature extraction for underwater acoustic target recognition," *Measurement*, Vol. 166, p. 108227, Jul. 2020, <https://doi.org/10.1016/j.measurement.2020.108227>
- [14] Z. Pu, Q. Zhang, Y. Xue, P. Zhu, and X. Cui, "A novel multi-feature fusion model based on pre-trained Wav2vec 2.0 for underwater acoustic target recognition," *Remote Sensing*, Vol. 16, No. 13, p. 2442, Jul. 2024, <https://doi.org/10.3390/rs16132442>
- [15] Y. Chang, X. Zhang, S. Song, Q. Song, and Y. Shen, "Rail crack defect recognition based on a multi-feature fusion algorithm using electromagnetic acoustic emission technique," *Measurement Science and Technology*, Vol. 34, No. 11, p. 115002, Nov. 2023, <https://doi.org/10.1088/1361-6501/ace840>
- [16] P. Qi, G. Yin, and L. Zhang, "Underwater acoustic target recognition using RCRNN and wavelet-auditory feature," *Multimedia Tools and Applications*, Vol. 83, No. 16, pp. 47295–47317, Oct. 2023, <https://doi.org/10.1007/s11042-023-17406-2>
- [17] B. Sun and X. Luo, "Underwater acoustic target recognition based on automatic feature and contrastive coding," *IET Radar, Sonar and Navigation*, Vol. 17, No. 8, pp. 1277–1285, Aug. 2023, <https://doi.org/10.1049/rsn2.12418>
- [18] Y. Sun et al., "Feature extraction methods for underwater acoustic target recognition of divers," *Sensors*, Vol. 24, No. 13, p. 4412, Jul. 2024, <https://doi.org/10.3390/s24134412>
- [19] S. Yao and T. Li, "Multi-domain acoustic feature fusion for speaker recognition," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pp. 1–6, 2024, <https://doi.org/10.1145/3696409.3700281>
- [20] H. Zhao, N. Huang, and H. Chen, "Knowledge enhancement for speech emotion recognition via multi-level acoustic feature," *Connection Science*, Vol. 36, No. 1, Dec. 2024, <https://doi.org/10.1080/09540091.2024.2312103>
- [21] P. Zhu et al., "SFC-Sup: robust two-stage underwater acoustic target recognition method based on supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, pp. 1–23, Nov. 2023, <https://doi.org/10.1109/tgrs.2023.3329653>
- [22] F.-M. Hoffmann, F. M. Fazi, E. G. Williams, and S. Fontana, "A general radiation model for sound fields and nearfield acoustical holography in wedge propagation spaces," *The Journal of the Acoustical Society of America*, Vol. 142, No. 3, pp. 1249–1260, Sep. 2017, <https://doi.org/10.1121/1.4998573>

- [23] K. Yuan et al., “Strong and agile wall-climbing robots capable of traversing obstacles via anisotropic acoustic adhesion,” *Research*, Vol. 9, Jan. 2026, <https://doi.org/10.34133/research.1038>
- [24] Y. Dun, S. Shan, and J. Liu, “Research on voiceprint recognition based on depth learning,” in *International Conference on Neural Networks, Information and Communication Engineering*, p. 93, 2021, <https://doi.org/10.1117/12.2615286>
- [25] S. Liu, C. Kim, and Y. Nam, “Speaker recognition based on machine learning classifiers for parent and child speaker diarization,” in *2024 International Conference on Platform Technology and Service (PlatCon)*, pp. 68–71, Aug. 2024, <https://doi.org/10.1109/platcon63925.2024.10830715>
- [26] K. Luo and L. Fu, “Research and application of voiceprint recognition based on a deep recurrent neural network,” in *Automatic control, mechatronics and industrial engineering*, CRC Press, 2019, pp. 309–316, <https://doi.org/10.1201/9780429468605-43>
- [27] Z. Soumaya, B. Drissi Taoufiq, N. Benayad, K. Yunus, and A. Abdelkrim, “The detection of Parkinson disease using the genetic algorithm and SVM classifier,” *Applied Acoustics*, Vol. 171, p. 107528, Jul. 2020, <https://doi.org/10.1016/j.apacoust.2020.107528>
- [28] Z. Zhao, Y. Wang, G. Shen, Y. Xu, and J. Zhang, “TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 3771–3782, 2023, <https://doi.org/10.1109/taslp.2023.3316458>
- [29] Y. Zhu, G. Li, S. Tang, R. Wang, H. Su, and C. Wang, “Acoustic signal-based fault detection of hydraulic piston pump using a particle swarm optimization enhancement CNN,” *Applied Acoustics*, Vol. 192, p. 108718, Mar. 2022, <https://doi.org/10.1016/j.apacoust.2022.108718>
- [30] Y. Zhu, G. Li, R. Wang, S. Tang, H. Su, and K. Cao, “Intelligent fault diagnosis of hydraulic piston pump combining improved LeNet-5 and PSO hyperparameter optimization,” *Applied Acoustics*, Vol. 183, p. 108336, Aug. 2021, <https://doi.org/10.1016/j.apacoust.2021.108336>
- [31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016, <https://doi.org/10.1109/cvpr.2016.90>
- [32] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017, <https://doi.org/10.1109/cvpr.2017.106>
- [33] X. Liu, H. Pan, and X. Li, “Object detection for rotated and densely arranged objects in aerial images using path aggregated feature pyramid networks,” in *MIPPR 2019: Pattern Recognition and Computer Vision*, p. 27, 2020, <https://doi.org/10.1117/12.2538090>
- [34] S. Feng, N. Zhang, and C. Zhao, “Path aggregation and dual supervision network for scene text detection,” in *Lecture Notes in Computer Science*, Vol. 12307, Cham: Springer International Publishing, 2020, pp. 66–77, https://doi.org/10.1007/978-3-030-60636-7_6
- [35] Z. Guo, H. Shuai, G. Liu, Y. Zhu, and W. Wang, “Multi-level feature fusion pyramid network for object detection,” *The Visual Computer*, Vol. 39, No. 9, pp. 4267–4277, Sep. 2022, <https://doi.org/10.1007/s00371-022-02589-w>
- [36] J.-F. Hu, J. Sun, Z. Lin, J.-H. Lai, W. Zeng, and W.-S. Zheng, “APANet: auto-path aggregation for future instance segmentation prediction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 7, pp. 1–1, Feb. 2021, <https://doi.org/10.1109/tpami.2021.3058679>



Zhixiang Zhang received the bachelor’s degree in mechanical engineering from North China Electric Power University in 2023. He is currently pursuing the Master’s degree at Xi’an Jiaotong University, Xi’an, Shaanxi, China. His current research interests include fault diagnosis.



Wenyu Gao received the bachelor’s degree in mechanical engineering from Ocean University of China in 2024. He is currently pursuing the master’s degree at Xi’an Jiaotong University, Xi’an, Shaanxi, China. His current research interests include fault diagnosis.



Yizhuo Gong received the bachelor's degree in mechanical engineering from Ocean University of China in 2024. He is currently pursuing the master's degree at Xi'an Jiaotong University, Xi'an, Shaanxi, China. His current research interests include fault diagnosis.



Yunhao Wang obtained a bachelor's degree in automation from Beijing Institute of Technology in 2023. He is currently pursuing a master's degree at Xi'an Jiaotong University in Xi'an, Shaanxi Province, China. His current research interests include reinforcement learning, embodied intelligence and VLA.



Zun Yin obtained a master's degree in materials science from the University of Science and Technology Beijing in 2009. He currently works at Xi'an Thermal Engineering Research Institute Co., Ltd. in Xi'an, Shaanxi Province, China. His current research interests include intelligent inspection, analysis and assessment of power station equipment.



Xiaomeng Li, a Senior Engineer, works at Huaneng Tongchuan Zhaojin Coal and Power Co., Ltd., where he specialises in steam turbine technology.



Damao Yang obtained a master's degree in automation from North China Electric Power University in 2006 and is currently employed at Huaneng Tongchuan Zhaojin Coal and Power Co., Ltd. His primary research focus is on coordinated control in thermal power stations.



Jin Zhou received the Ph.D. degree from the University of Liverpool, Liverpool, U.K. He is currently a National Youth Distinguished Professor and Doctoral Supervisor with Xi'an Jiaotong University, Xi'an, Shaanxi, China, and an Honorary Research Fellow with the University of Liverpool, Liverpool, U.K. His current research interests include control, dynamics and fault diagnosis.



Jun Peng received the Ph.D. degree in mechanical engineering from the University of Wisconsin-Madison, Madison, WI, USA. He is currently a Professor with Xi'an Jiaotong University, Xi'an, Shaanxi, China. His current research interests include control, dynamics and fault diagnosis.