Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 20, No. 1, 2026

Fault Diagnosis Method for Pumping Units with Small Samples Based on Improved ConvNeXt and Transfer Learning

Ze Li

School of Mechanical Engineering, Xi'an Shiyou University, Xi'an, Shaanxi, 710065, China

Abstract: The scarcity of real fault samples in pumping units degrades the generalization of deep learning models. An improved ConvNeXt diagnosis method integrating physical feature perception and hierarchical transfer learning is proposed. First, according to the "low texture, high geometric structure" of dynamometer cards, ConvNeXt V2 is lightweight adapted and embedded with a channel attention module (CAM). By enhancing sensitivity to feature channels representing displacement and load variations, an adapted model with 22.1M parameters is constructed. Second, a two-stage "mechanism-driven pre-training and measured hierarchical fine-tuning" strategy is designed: simulation dynamometer cards generated by a dynamic model inject physical priors; during fine-tuning, shallow general features are frozen, deep semantic layers are differentially fine-tuned, and the classification head is retrained, achieving efficient cross-domain knowledge transfer. Experiments on a measured dataset containing eight working conditions show that the proposed method achieves an accuracy of 96.12% and a macro-average F1 score of 95.84%, with convergence speed approximately 30% faster than mainstream models. Ablation experiments confirm the synergistic effect of the physical perception structure and hierarchical transfer; for the extremely small-sample "sand sticking" fault, the F1 score significantly increases from 88.5% to 93.8%. Visualization analysis reveals the response mechanism of the CAM module to physical signals such as load surges, providing a high-confidence solution for industrial intelligent maintenance under small samples.

Keywords: Fault Diagnosis; Pumping Unit; Dynamometer Card; Few-shot Learning; ConvNeXt; Transfer Learning; Channel Attention.

1. Introduction

1.1. Background and Significance

Beam pumping units are the core oil extraction equipment in onshore oilfields, and their operating status directly determines the efficiency of crude oil recovery. The dynamometer card, a closed curve of polished rod load versus displacement, [1]. Traditional manual diagnosis relies on experts' empirical interpretation of geometric features of dynamometer cards (area, slope, concavity/convexity), [2]. The construction of digital oilfields urgently requires real-time online monitoring of massive oil wells, making the development of intelligent fault diagnosis technology imperative [3].

1.2. Related Work

Deep learning has made significant progress in the field of mechanical fault diagnosis. Convolutional neural networks (CNNs), due to their powerful automatic spatial feature extraction capability, have been widely used for dynamometer card recognition. Early work introduced classic visual networks (such as AlexNet) for dynamometer card classification, [5] . One-dimensional CNNs for extracting sequential features, as well as combinations of CNNs with recurrent/attention modules to capture temporal-spatial correlations, [21] . However, these methods generally rely on large-scale, balanced, and well-annotated training datasets. In real oilfields, there is a serious shortage of fault samples, leading data-driven models prone to overfitting and poor generalization.

Transfer learning has been introduced to overcome the data scarcity bottleneck, with the core idea of transferring knowledge from a source domain (e.g., public datasets, simulation data) to a target domain (e.g., measured data from specific equipment). Simulated dynamometer cards can inject physical priors and serve as rich source domain data to assist training [9], [8]. However, directly transferring natural image pre-trained models or simple mixed training often neglects the domain shift between simulation and reality, which may cause the model to learn domain-specific spurious features, leading to catastrophic forgetting or overfitting [9], [12].

Moreover, the development of modern vision networks provides new architectural choices for dynamometer card diagnosis. The ConvNeXt family modernizes CNNs by borrowing design ideas from Vision Transformers (e.g., larger kernels, inverted bottlenecks, LayerNorm) and achieves excellent performance on natural image tasks ; ConvNeXt V2 further improves representation ability through co-design of network architecture and self-supervised pre-training (e.g., Masked Autoencoders) [12]. However, these general vision architectures are not specifically designed for dynamometer cards ("low texture, high geometric structure"), and direct transfer may contain considerable parameter redundancy irrelevant to dynamometer card diagnosis, as well as sensitivity to irrelevant textures [10].

Channel attention mechanisms adaptively recalibrate feature channels to guide the model to focus on task-relevant channels, and have proven effective in various vision tasks and industrial diagnosis [3], [4], [19]. Combining simulation-driven transfer learning with physically-aware model structure design is expected to improve diagnostic performance and interpretability under small-sample conditions.

1.3. Main Contributions

To address the above issues, this paper proposes an improved ConvNeXt method integrating physical feature perception and hierarchical transfer learning. The main contributions are as follows:

(1) Model structure adaptation and physical feature perception: For the first time, ConvNeXt V2 is introduced into pumping unit dynamometer card fault diagnosis, and a channel attention module (CAM) is embedded to make the model adaptively enhance feature channels related to displacement and load variations, constructing a lightweight (22.1M parameters), efficient, and interpretable dedicated adapted model.

(2) Hierarchical transfer learning strategy: A "mechanism-driven pre-training and measured hierarchical fine-tuning" framework is proposed. Simulation data generated by a dynamic model inject physical priors during pre-training; during fine-tuning, a hierarchical parameter update mechanism (freezing the first three stages, fine-tuning the last stage, and retraining the classification head) is designed to achieve efficient and precise knowledge transfer from the simulation domain to the measured domain.

(3) Comprehensive experiments and interpretability analysis: Comparative and ablation experiments are designed on a real pumping unit dynamometer card dataset. Visualization of channel attention weights reveals the physical basis of model decisions, providing a high-confidence case reference for small-sample industrial fault diagnosis.

2. Related Theories and Methods

2.1. Principle of Dynamometer Card Fault Diagnosis

The dynamometer card is a closed curve of load versus displacement during one stroke cycle of the pumping unit polished rod. Different fault types exhibit specific geometric shapes on the dynamometer card: rod break appears as a "knife handle" shape, gas influence as an "obese" shape, sand sticking as a "sawtooth" shape, and pump leakage as a "notch" [2]. These morphological features directly correspond to downhole physical processes, providing a theoretical basis for image-based fault diagnosis.

2.2. ConvNeXt Network Architecture

ConvNeXt modernizes traditional CNNs by systematically borrowing design concepts from Vision Transformers [2]. Core components include:

Larger kernels: 7×7 depthwise convolutions to enlarge receptive fields.

Inverted bottleneck structure: Following the MLP block in Transformers, adopting an "expand-compress" design.

LayerNorm instead of BatchNorm: Improving training stability.

Fewer activation functions: Reducing the use of GELU to simplify the structure.

ConvNeXt V2 further introduces Global Response Normalization (GRN) and co-designs with self-supervised pre-training (e.g., Masked Autoencoders) to enhance representation capability .

2.3. Channel Attention Mechanism

Channel attention mechanisms model dependencies between feature channels to adaptively recalibrate channel importance. SENet first proposed generating channel weights via global average pooling and fully connected layers [4]; CBAM combines channel and spatial attention to strengthen feature representation [11]. A comprehensive review of attention mechanisms can be found in [19]. The CAM module used in this paper combines global average pooling and global max pooling, followed by a shared MLP to model nonlinear dependencies between channels, capturing feature responses more comprehensively.

2.4. Transfer Learning and Few-shot Learning

Transfer learning aims to transfer knowledge from a source domain to a target domain to solve the problem of scarce target domain samples [7]. In mechanical fault diagnosis, common methods include fine-tuning pre-trained models, domain adaptation, and meta-learning [18]. Few-shot learning specifically refers to scenarios where each class in the target domain has only a few labeled samples, imposing higher requirements on model generalization. In recent years, using simulation data for pre-training combined with hierarchical fine-tuning or structural adaptation has been proven to significantly improve target domain performance in fault diagnosis [9].

3. Improved ConvNeXt Fault Diagnosis Model

3.1. Data Preprocessing and Augmentation

The original input of the model is a pair of displacement-load time series data collected during one stroke cycle of the pumping unit polished rod. The two one-dimensional sequences are mapped onto a two-dimensional plane to form a closed dynamometer card contour, and then uniformly scaled to a 64×64 pixel single-channel grayscale image. This resolution balances feature retention and computational efficiency: it is sufficient to preserve key morphological features of various faults (e.g., "knife handle" rod break, "obese" gas lock, "sawtooth" sand sticking), while reducing subsequent convolution memory usage and inference time.

To improve model robustness under limited data and avoid overfitting, lightweight data augmentation is applied to the input images during training:

Random horizontal flip (probability 0.5): simulates symmetry of stroke direction.

Small-angle random rotation (±5°): simulates sensor installation tilt error (large rotations that would destroy the physical coordinate meaning are avoided).

Random brightness and contrast adjustment (±10%): simulates different lighting or sensor gain conditions.

Augmentation increases data diversity without changing the essential nature of the faults.

3.2. ConvNeXt V2 Backbone Adaptation and Lightweight Design

ConvNeXt-Tiny was originally designed for massive high-texture natural images, with generous channel settings at each stage. Dynamometer cards have sparse texture but critical geometric structure; direct use of the original model leads to parameter redundancy and may cause the network to over-focus on irrelevant texture details. In this paper, based on ConvNeXt V2 [11], the network structure is simplified for dynamometer card characteristics: shallow stages (Stage 1–2) use fewer channels to extract basic low-level features (edges,

gradients); deep stages (Stage 3–4) appropriately increase channels to fuse complex fault structure patterns. The optimized configuration of channel number *C* and number of ConvNeXt blocks *B* at each stage is shown in Table 1. The parameter count is reduced from 28.6M in the original ConvNeXt-Tiny to 22.1M, and FLOPs are correspondingly reduced, achieving lightweight design suitable for few-shot learning scenarios. The overall architecture of the adapted network is shown in Fig. 1.

Table 1. Optimized configuration of ConvNeXt model for dynamometer card diagnosis

Stage	Feature map size	Channels C	ConvNeXt blocks B	Main learning features
Stage 1	64×64	64	3	Basic edges, corners, gradient changes
Stage 2	32×32	128	3	Abstract contours, simple geometric combinations
Stage 3	16×16	256	6	Complex fault structure patterns, inter-region relationships
Stage 4	8×8	512	3	Global semantic features, class discrimination information

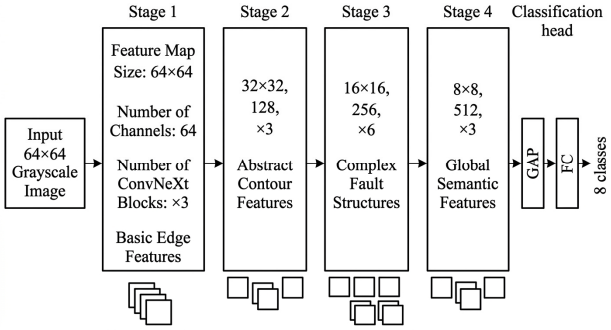
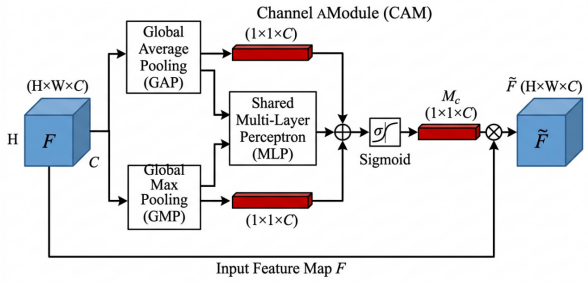


Fig 1. Overall architecture of the adapted ConvNeXt V2

3.3. Physical Feature Perception Channel Attention Mechanism

Different channels of the high-dimensional feature maps of dynamometer cards respectively represent geometric attributes such as slope, intercept, and local curvature, which are highly related to physical information such as load fluctuation and displacement exceeding limits. To make the model adaptively focus on key physical features, a channel attention module (CAM) is embedded in the ConvNeXt blocks, as shown in Fig. 2. The module first aggregates spatial information via global average pooling (GAP) and global max pooling (GMP), then models nonlinear dependencies between channels through a shared multi-layer perceptron (MLP), generating channel attention weights as follows:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$



Formula (1): $M_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F)))$

Fig 2. Structure of channel attention module (CAM)

where F is the input feature map and σσ is the Sigmoid function. Through this mechanism, the model dynamically strengthens feature channels that capture load drastic changes (e.g., sand sticking, piston sticking) or displacement anomalies (e.g., rod break), while suppressing background noise and redundant texture channels, achieving physical meaning reweighting at the feature level. The CAM design is inspired by methods such as SENet and CBAM, with a trade-off in parameter and computational overhead [12].

3.4. Weighted Cross-entropy Loss Function

Measured fault data exhibit class imbalance: as shown in Table 2, "waxing" and "insufficient supply" have more samples, while "sand sticking" samples are scarce. Standard cross-entropy loss would bias towards the majority classes, leading to poor recognition of minority classes. A class-weighted cross-entropy loss function is introduced:

$$L_W CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \log(\hat{y}_{i,c})$$

where N is the batch size, $C = 8$ is the number of fault classes, y_i, c is the indicator function (1 if sample i belongs to class c, otherwise 0), $\hat{y}\_i, c$ is the predicted probability, and w_c is the class weight. Inverse frequency weighting is adopted, i.e., $w\_c = N\_max / N\_c$, where N_max is the sample count of the majority class and N_c is the sample count of class c. This mechanism assigns larger gradients to the loss of rare fault samples, forcing the model to treat all classes equally.

3.5. Hierarchical Transfer Learning Framework

The hierarchical transfer learning framework adopted in this paper consists of two stages, as illustrated in Fig. 3.

Stage 1: Source domain (simulation) pre-training. A large-scale and diverse simulation dynamometer card dataset is generated using a pumping unit system dynamics simulation software based on the Gibbs one-dimensional wave equation [16], covering all studied fault types. This simulation dataset is used to fully pre-train the adapted ConvNeXt V2 model constructed in Section 3.2, enabling it to learn general morphological variation patterns and physical prior knowledge of dynamometer cards under various theoretical fault conditions.

Stage 2: Target domain (measured) hierarchical fine-

tuning. The pre-trained model parameters are loaded into the target diagnostic task. To avoid overfitting caused by fine-tuning all parameters on a small amount of measured data while retaining pre-trained knowledge, a hierarchical parameter update strategy is designed:

Freeze parameters of Stage 1, Stage 2, and Stage 3: These layers learn general features such as edges and contours, which are highly transferable between simulation and measurement. Freezing them preserves general knowledge and reduces the number of trainable parameters (by about 70%), preventing overfitting.

Fine-tune parameters of Stage 4: The deep layer is responsible for learning abstract fault semantic features. It is finely adjusted with a low learning rate (1e-5) to adapt to the measured data distribution.

Replace and retrain the classification head: The original 1000-class classification head is replaced with a new randomly initialized fully connected layer (outputting 8 classes), which is trained from scratch to map the fine-tuned features to specific fault categories.

This hierarchical strategy achieves progressive knowledge transfer from general geometric features to specific fault semantics, effectively mitigating domain shift and maximizing the value of limited measured samples.

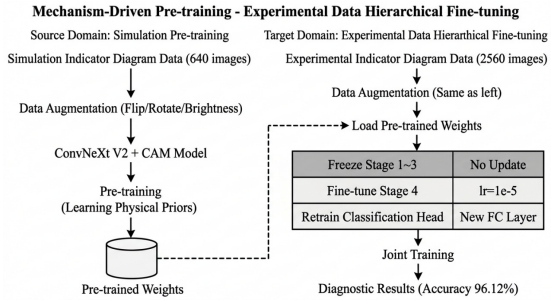


Fig 3. Framework of two-stage transfer learning

4. Experiments and Results Analysis

4.1. Experimental Setup

4.1.1. Dataset Description

The data come from actual production blocks of a domestic oilfield, comprising 8 working conditions: normal condition and 7 typical fault types (rod break, pump leakage, gas lock (gas influence), waxing, sand sticking, piston sticking, and insufficient supply). The total number of samples is 3200, of which 2560 are field measured data and 640 are simulation data generated by simulation software based on the Gibbs wave equation. The sample distribution for each category is shown in Table 2. The measured data are significantly imbalanced. All data are randomly divided into training, validation, and test sets in a 6:2:2 ratio, ensuring consistent category proportions. The training set is augmented with the data enhancement described in Section 3.1, while the validation and test sets are only standardized.

Table 2. Composition of the pumping unit dynamometer card dataset

Working condition	Measured samples	Simulation samples
Normal	310	80
Rod break	295	60
Pump leakage	270	70
Gas lock	280	70
Sand sticking	260	60
Waxing	430	150
Piston sticking	270	60
Insufficient supply	445	90

4.1.2. Comparison Models and Evaluation Metrics

ResNet-50 [5], EfficientNet-B0 [6], and original ConvNeXt-Tiny [2] are selected as baseline models. All models are implemented in PyTorch and trained on an NVIDIA RTX 3090 with the AdamW optimizer, batch size 32, and the same learning rate scheduling strategy. Baseline models are initialized with ImageNet-1K pre-trained weights and fine-tuned on the training set. Training parameters for the proposed method: AdamW optimizer, initial learning rate 1×10^{-4} (fine-tuning stage for Stage 4: 1×10^{-5}), weight decay 1×10^{-4}, batch size 32, maximum epochs 100, early stopping (stop when validation loss does not decrease for 5 consecutive epochs).

Evaluation metrics include accuracy, macro-average F1 score, precision/recall/F1 for each class, convergence epoch (epoch at early stopping), and FLOPs. In addition, to measure the actual improvement on small-sample classes (e.g., sand sticking), a dedicated few-shot F1 comparison is included in the ablation experiments.

4.2. Performance Comparison with Baseline Models

Table 3 shows the performance comparison between the proposed method (Ours) and the baseline models on the test set. Figure 4 displays the training curves of accuracy and loss. The proposed method outperforms all baselines in both accuracy and convergence speed.

Table 3. Performance comparison of different diagnosis models on the test set

Model	Parameters (M)	FLOPs (G)	Accuracy (%)	Macro F1 (%)	Convergence epochs
Ours (adapted ConvNeXt V2)	**22.1**	**4.2**	**96.12**	**95.84**	**32**
ConvNeXt-Tiny (original)	28.6	4.5	94.78	94.33	45
ResNet-50	25.6	4.1	95.05	94.62	50
EfficientNet-B0	5.3	0.4	93.44	92.87	38

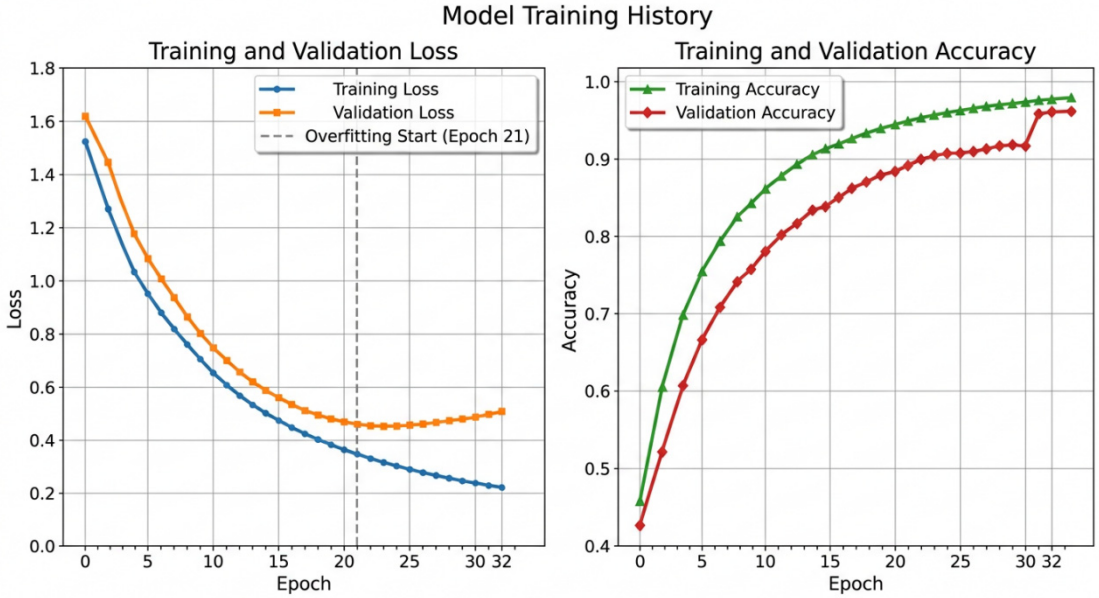


Fig 4. Training curves of accuracy and loss

The proposed method achieves an accuracy of 96.12% and a macro F1 of 95.84%, leading all baselines. Compared to the original ConvNeXt-Tiny, accuracy improves by 1.34 percentage points, macro F1 by 1.51 points, while parameters are reduced by 22.7% and FLOPs by 6.7%. Compared to ResNet-50, accuracy improves by 1.07 points and convergence epochs decrease by 36%. EfficientNet-B0 has the smallest parameter count but the lowest diagnostic accuracy, indicating that its extreme lightweight design may compromise the representation of dynamometer card geometric structures [14]. The training curves show that with simulation pre-training and hierarchical fine-tuning, the proposed model trains more stably, validation loss converges faster, and convergence epochs are significantly reduced (≈30%), verifying the synergy of the transfer strategy and structural adaptation.

4.3. Ablation Experiments

To quantify the contributions of the core strategies, four configurations are designed for ablation experiments, with results shown in Table 4.

Comparing Exp-A and Exp-B, transfer learning improves accuracy by 3.25 percentage points and sand sticking F1 by 5.3 points, demonstrating the necessity of injecting physical priors via simulation pre-training[16]. Comparing Exp-A and Exp-C, hierarchical fine-tuning outperforms simple mixed training by 1.09 points in accuracy, indicating that the hierarchical strategy more effectively handles domain shift: freezing shallow layers preserves simulation physical priors while fine-tuning only the deep layers adapts to the measured distribution, avoiding feature confusion or catastrophic forgetting [17]. Comparing Exp-A and Exp-D, model adaptation (channel reduction + CAM) brings an accuracy improvement of 1.59 points under the same transfer strategy, proving the necessity of structural optimization (i.e., ConvNeXt adjustment for dynamometer cards and attention embedding). For the sand sticking fault with the fewest samples, the complete method achieves an F1 of 93.8%, an increase of 5.3 points over using only measured data, indicating that CAM's reinforcement of load-related channels enables the model to accurately capture the core physical features of sand sticking[18].

Table 4. Ablation study on model performance under different training strategies

Exp.	Model structure	Training strategy	Accuracy (%)	Macro F1 (%)	Sand sticking F1 (%)
Exp-A	**Adapted ConvNeXt V2 + CAM (Ours)**	**Simulation pre-training + measured hierarchical fine-tuning**	**96.12**	**95.84**	**93.8**
Exp-B	Adapted ConvNeXt V2 + CAM	Only measured data (random init)	92.87	92.24	88.5
Exp-C	Adapted ConvNeXt V2 + CAM	Simple mixed training (simulation + measured)	95.03	94.58	92.3
Exp-D	Original ConvNeXt-Tiny	Simulation pre-training + measured hierarchical fine-tuning	94.53	94.02	91.6

4.4. Interpretability Analysis

To verify the physical consistency of model decisions, the attention weights of the CAM module under different fault inputs are extracted and aggregated according to physical attributes (load-mutation-related channels and displacement-anomaly-related channels). The results are shown in Fig. 5.

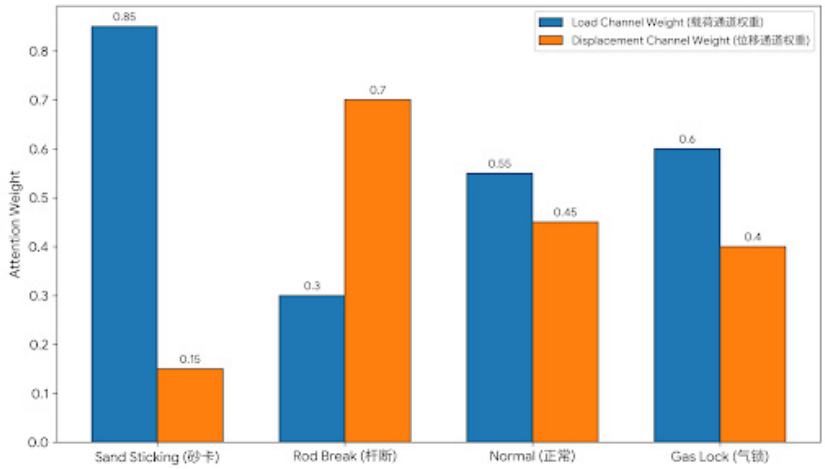


Fig 5. Distribution of channel attention weights under different fault types

Visualization shows that under sand sticking fault, load-related channels have a weight of about 0.85, while displacement-related channels are only 0.15, consistent with the physical mechanism of sand sticking (drastic load change with almost no displacement change). Under rod break fault, displacement-related channel weight rises to about 0.7, and load-related channels are about 0.3, corresponding to the physical feature of abnormal displacement fluctuation after rod break. Under normal condition, the weights of the two types of channels are close (≈0.55/0.45), reflecting the good coupling between displacement and load. Under gas lock fault, load channels are slightly higher (≈0.6/0.4), consistent with intensified load fluctuation due to gas influence. This alignment of "visual morphology → physical meaning" verifies that the model's discriminative basis under small-sample conditions is consistent with expert experience, thereby improving the credibility of diagnostic results [19].

5. Conclusion and Future Work

5.1. Conclusion

This paper addresses the small-sample problem in pumping unit fault diagnosis by proposing an improved ConvNeXt method integrating physical feature perception and hierarchical transfer learning. The main conclusions are as follows:

(1) Lightweight model adaptation: For the "low texture, high geometric structure" characteristics of dynamometer cards, ConvNeXt V2 is channel-reduced and depth-optimized, reducing parameters to 22.1M and FLOPs to 4.2G. The embedded CAM module enables the model to adaptively enhance feature channels related to load mutations and displacement anomalies, improving feature discrimination and interpretability.

(2) Hierarchical transfer strategy: The "mechanism-driven pre-training and measured hierarchical fine-tuning"

framework is proposed. By freezing the first three stages, fine-tuning the last stage, and retraining the classification head, efficient transfer of simulation knowledge to measured data is achieved, effectively alleviating small-sample overfitting. Experiments on eight working conditions show that the proposed method achieves an accuracy of 96.12% and a macro F1 of 95.84%, with convergence speed about 30% faster than mainstream models.

(3) Synergy effect verification: Ablation experiments confirm the synergistic effect of the physical perception structure and hierarchical transfer; for the sand sticking fault with the fewest samples, the F1 score increases by 5.3 percentage points. Visualization of channel attention weights reveals the model's response mechanism to physical signals such as load surges, confirming that the model learns physical knowledge consistent with expert experience.

(4) Industrial application value: The proposed method has low parameter count and computational cost, making it suitable for edge deployment. Visualization analysis confirms the consistency between model decisions and mechanical mechanisms, providing interpretable support for trustworthy intelligent maintenance in industrial scenarios.

5.2. Future Work

(1) Domain adaptation enhancement: Explore unsupervised domain adaptation techniques such as adversarial training to further reduce the distribution gap between simulation and measured data and improve transfer efficiency [13].

(2) Multi-fault coupling diagnosis: The current model focuses on single-label classification, but real working conditions may involve multiple concurrent or coupled faults. Future work can study multi-label classification or fault decoupling methods [14].

(3) Model lightweight deployment: Advance lightweight techniques such as pruning and quantization to adapt to the resource constraints of oilfield edge computing devices, enabling real-time, low-power diagnostic applications [6].

References

[1] J. H. Lin, Y. Zhou, T. Ding, et al., "Fault classification evaluation method of pumping wells based on improved EfficientNet," Science Technology and Engineering, 2023. [Online]. Availablehttps://kns.cnki. net/kcms/ detail/ detail. aspx?dbcode=CJFD&dbname=CJFDLAST2023&filename=KJYJ202311029.

[2] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976–11986. doi: 10.1109/CVPR52688.2022.01594.

[3] S. W. Pan and S. J. Yang, "An overview of transfer learning in deep learning," IEEE Trans. Cybern., 2021. doi: 10.1109/ TCYB. 2021.3106516.

[4] R. S. Wang and Y. F. Fu, "Fault diagnosis of rotating machinery using a new deep learning method with data augmentation," Measurement, 2018. doi: 10.1016/j. measurement. 2017.11.020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778. doi: 10.1109/ CVPR. 2016.90.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114. [Online]. Available: https: // proceedings.mlr.press/v97/tan19a.html.

[7] S. J. Pan and Q. Yang, "A comprehensive review on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010. doi: 10.1109/TKDE.2009.191.

[8] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, and R. Chen, "Deep transfer learning for fault diagnosis with a small number of samples," Neurocomputing, 2020. doi: 10.1016/j. neucom. 2019. 09.088.

[9] J. Sun, C. Ding, and S. Gong, "Domain adaptation for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., 2016. doi: 10.1109/TPAMI.2016.2570081.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[11] J. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19. doi: 10.1007/978-3-030-01237-3_1.

[12] X. Zhai, A. Kolesnikov, L. Beyer, et al., "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," arXiv preprint arXiv:2301.00808, 2023.

[13] J. Lv, W. Li, S. Wu, and M. Zhang, "Deep domain adaptation: A survey of the state of the art," Neural Networks, 2019. doi: 10.1016/j.neunet.2019.04.004.

[14] T. Y. Chen, D. M. Zhang, and X. W. Sun, "A review of few-shot fault diagnosis methods for rotating machinery based on deep learning," Journal of Vibration and Shock, 2023. [Online]. Available: https://www.sciopen.com/article/10.19693/j.issn.1673-3185.04175.

[15] X. Ding, H. Mao, H. Zhang, Q. Xie, and X. Zhang, "Improving transferability of convolutional neural networks via structural re-parameterization," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2021. [Online].Available: https:// proceedings. neurips.cc/paper/2021/hash/c8d5040e34f664a78086064f51e7ca30-Abstract.html.

[16] Y. F. Li, P. Y. Song, C. M. Zheng, et al., "Production prediction of ultra-high water-cut oil wells based on convolutional neural network and transfer learning," Petroleum Geology and Recovery Efficiency, 2020. [Online]. Available: https:// kns. cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2020&filename=YQDC202005018.

[17] P. Ren, Z. Xu, Z. Luo, and W. Liu, "Deep learning based multi-fault diagnosis of rolling bearings using 1D convolutional neural network," Measurement, 2019. doi: 10.1016/j. measurement. 2018.10.046.

[18] J. Snell, K. Swersky, and R. S. Zemel, "Meta-learning for few-shot image classification," in Int. Conf. Learn. Representations (ICLR).

[19] J. Zhang, J. Hu, Y. Wang, and L. Yang, "A survey on attention mechanisms in deep learning," Pattern Recognit., 2020. doi: 10.1016/j.patcog.2020.107402.

[20] H. Li, X. Liu, M. Cao, and X. Li, "Transfer learning with deep convolutional neural networks for fault diagnosis of rolling bearings," Appl. Soft Comput., 2019. doi: 10.1016/j. asoc. 2019. 105470.

[21] X. G. Li, Z. G. Wang, and L. Q. Sun, "Review of transfer learning methods for rolling bearing faults under variable working conditions and small samples," Machinery & Electronics, 2024. [Online]. Available: http://www. stae.com. cn/ jsygc/article/issue/2024_0_10.