

Original Paper

Multi-task ensemble learning for real-time efficiency prediction and condition diagnosis in beam-pumping well systems

Biao Ma, Shi-Min Dong^{*}

School of Mechanical Engineering, Yanshan University, Qinhuangdao, 066004, Hebei, China

ARTICLE INFO

Article history:

Received 26 October 2025

Received in revised form

20 March 2026

Accepted 29 June 2026

Available online 3 July 2026

Edited by Min Li

Keywords:

Multi-task learning

Ensemble learning

Real-time prediction

Condition diagnosis

ABSTRACT

Currently, real-time efficiency prediction and condition diagnosis for beam-pumping well systems predominantly rely on single-task models. However, conventional single-task approaches typically require constructing two separate models, resulting in doubled computational resource demands and an inability to effectively exploit the coupling relationships between prediction and diagnosis tasks. To address these limitations, this paper proposes a series of multi-task learning-based models for real-time efficiency prediction and condition diagnosis of beam-pumping well systems. First, we propose a multi-task learning model based on multimodal dual-path fusion, which enables effective feature sharing between efficiency prediction and condition diagnosis tasks using a dual-path network architecture. Second, we introduce a multimodal cross-scale fusion multi-task learning model designed to effectively capture the coupling relationships among multimodal features across different scales. Finally, to further improve the prediction accuracy of system efficiency and enhance the precision of condition diagnosis, we develop a meta-learning-based dynamic voting ensemble multi-task learning model, which dynamically integrates predictions from multiple tasks. To verify the effectiveness of the proposed methods, comprehensive experiments and ablation studies are conducted on a dataset consisting of 3938 actual beam-pumping wells. Experimental results demonstrate that, compared with the baseline networks, the proposed method achieves significant performance improvements: relative to the comparison networks, it reduces RMSE by 18.91%–44.73%, increases R^2 by 13.03%–101.13%, and enhances Accuracy and F1-score by 2.20%–32.84% and 2.16%–34.31%, respectively. Compared with the multimodal cross-scale multi-task learning network, it decreases RMSE by 6.86%–32.69%, raises R^2 by 3.66%–38.61%, and improves Accuracy and F1-score by 2.30%–173.35% and 2.31%–293.73%, respectively. Relative to the multi-task learning-based on meta-learning dynamic voting ensemble, it reduces RMSE by 17.30%–36.80%, increases R^2 by 5.40%–20.00%, and boosts Accuracy and F1-score by 1.77%–5.92% and 1.76%–6.17%, respectively. When compared with the ablation networks, the baseline network achieves RMSE reductions of 6.05%–27.78%, R^2 improvements of 3.05%–25.51%, and Accuracy and F1-score enhancements of 4.59%–53.73% and 7.76%–65.18%, respectively. Furthermore, the network exhibits significant sensitivity to hyperparameters, with notable performance fluctuations observed under different hyperparameter configurations.

© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As a key oil-extraction device in onshore oilfields, the system efficiency and operating conditions of beam-pumping wells have a direct impact on crude oil output and production costs. In the

context of digital oilfield development, accurately predicting system efficiency in real time while simultaneously diagnosing abnormal operating conditions is essential for intelligent parameter adjustment and energy optimization. However, real-world field environments are often complex, characterized by high noise, multivariable coupling, and a scarcity of fault samples. Traditional monitoring methods, which rely on manual expertise or single physical models, are insufficient to meet the requirements for real-time performance and robust fault tolerance. Therefore, there is a pressing need for a multi-task learning model

^{*} Corresponding author.

E-mail address: ysudshm@ysu.edu.cn (S.-M. Dong).

Peer review under the responsibility of China University of Petroleum (Beijing).

that can simultaneously perform system efficiency prediction and operating condition diagnosis for beam-pumping wells.

Currently, system efficiency prediction for beam-pumping wells primarily falls into two categories: mechanism-based models and data-driven models using historical data mining. Mechanism-based models typically involve complex mathematical formulations. Although these models incorporate subsurface fluid dynamics and the surface mechanical properties of the pumping unit—offering valuable theoretical insights into beam-pumping system mechanics—they have several critical limitations. First, their implementation demands extensive prior data and involves mathematically intensive computations. Second, they struggle to capture the system's inherent nonlinear dynamics and time-varying characteristics. Most importantly, the substantial computational overhead of these models severely limits their feasibility for real-time prediction and adaptive control. In contrast, data-driven models based on historical records often employ simple neural networks for pattern discovery. However, such methods generally lack the responsiveness necessary for real-time applications. Currently, fault diagnosis for beam-pumping wells primarily relies on two types of models: those based on historical data mining and those utilizing electric power curve analysis. Models based on historical data rely exclusively on past records, which limits their applicability in real-time scenarios. Although power curve-based models can improve diagnostic timeliness and reduce operational costs, studies have shown that depending solely on power curve information is often insufficient for accurate fault identification. With continuous advancements in information technology and sensor technologies, a wide range of multimodal data sources has become available, creating new opportunities to enhance diagnostic accuracy. Therefore, it is imperative to develop condition diagnosis models for beam-pumping wells that leverage multimodal feature fusion. Both the efficiency-prediction and condition-diagnosis models for beam-pumping wells currently rely on separate single-task frameworks. This forces redundant processing of identical sensor data, ignores the underlying physical correlations between tasks, and prevents sharing of diagnostic features—reducing feature discriminability. Moreover, running two independent models doubles the computational burden on edge devices, hindering real-time responsiveness. Therefore, it is essential to develop a unified multi-task learning model that can simultaneously predict system efficiency and diagnose operating conditions in beam-pumping wells.

This paper presents three multi-task learning models for real-time efficiency prediction and condition diagnosis in beam-pumping wells. The first model employs multimodal dual-path fusion to jointly predict efficiency and detect operating anomalies. The second leverages multimodal cross-scale fusion to capture feature interactions across different resolutions. The third integrates a meta-learning based dynamic voting ensemble to further enhance efficiency prediction accuracy and condition diagnosis performance. We validate the proposed models through both full-scale and ablation experiments on a dataset of 3938 real-world wells, demonstrating substantial improvements in both efficiency forecasting and condition diagnosis. The main contributions of this work can be summarized as follows:

- (1) We propose a multimodal dual-path fusion multi-task learning model for real-time system efficiency prediction and condition diagnosis.
- (2) We introduce a multimodal cross-scale fusion multi-task learning model to capture coupling relationships among features at different resolutions for real-time efficiency prediction and condition diagnosis.

- (3) We develop a meta-learning-driven dynamic voting ensemble multi-task learning model to further enhance system efficiency prediction performance and condition diagnosis accuracy.
- (4) We conducted a comprehensive experimental analysis to determine the effectiveness of the proposed method.

2. Related work

2.1. Predictive modeling of pumping well system efficiency

Currently, system efficiency prediction models for beam-pumping wells can be classified into mechanism-based approaches and historical data-driven approaches. Mechanism-based models originate from studies on the longitudinal vibration of the sucker rod string. Gibbs (1963) first formulated the longitudinal vibration equation of the sucker rod string. Lekia and Evans (1965) introduced the rod–fluid coupled vibration equation, and Lukasiewicz (1991) extended the model to deviated wells. Wang et al. (2023) analyzed the impact of gas–liquid separation on rod-string vibration, while Xing and Zhou (2015) and Xing (2019) proposed nonlinear longitudinal vibration equations. Moreno and Garriz (2020) developed a nonlinear vibration model for directional wells, and Yin et al. (2021) provided an analytical solution for predicting the behavior of multi-cone pumping systems. Wang and Dong (2020) then investigated longitudinal vibration under down-hole track excitation in curved wells. Li et al. (2019) proposed a vibration equation based on a novel boundary model. Tarmigh et al. (2023) coupled transient multiphase flow equations with rod deformation equations to enhance numerical analysis. Further developments include Wang's thermo-mechanical coupling vibration model (Wang and Lv, 2021), Li's consideration of viscous and local damping losses (Li et al., 2019), and Ma's multiphase flow-based longitudinal vibration equation (Ma and Dong, 2023). Although mechanism-based efficiency prediction methods for beam-pumping wells integrate subsurface fluid dynamics with surface mechanical characteristics—providing valuable theoretical insights into well dynamics—they exhibit critical limitations. First, model development typically relies on simplifying assumptions and entails complex mathematical formulations. Second, these methods are limited in capturing the system's inherent nonlinear dynamics and time-varying behavior. To address these shortcomings, some researchers have turned to data-driven models based on historical records. Tan et al. (2022) proposed a timeseries-based efficiency prediction model, while Ma and Dong (2024a,b) introduced both a graph neural network-based predictor and a stacking ensemble learning approach for efficiency forecasting. However, such methods depend on extensive, high-quality historical datasets, resulting in inadequate real-time responsiveness, relatively low prediction accuracy, and limited robustness (Wang et al., 2019, b; Ju et al., 2018). Therefore, it is necessary to develop beam-pumping well efficiency prediction models with enhanced real-time performance.

2.2. Diagnostic model of pumping well conditions

Extensive work has addressed condition diagnosis for beam-pumping wells. Li et al. (2023) first proposed a multi-feature fusion fault-diagnosis model. Zheng and Gao (2017) employed an optimized continuous hidden Markov model. Feng et al. (2025) introduced a torque-indicator chart-based approach, while Wei and Gao (2020) presented a deep-and-wide learning system using motor-power data. Wang et al. (2025) proposed a parallel-learning-based augmentation mechanism, and Sreenivasan and Krishna applied transfer learning for diagnosis (Sreenivasan and

Krishna, 2024). Despite these advances, current models remain either purely historical-data driven or rely solely on electric-power parameters. Historical-data approaches lack real-time responsiveness and incur high diagnostic costs, whereas power-parameter-based methods struggle with complex operating conditions. Therefore, it is essential to develop feature-enhanced, multimodal diagnosis models for beam-pumping wells.

2.3. Multi-task learning model

Multi-task learning (MTL) has achieved remarkable success across diverse domains. Yang et al. (2025) introduced an advanced multi-task deep neural network for organic toxicity prediction. Zhang et al. (2025) developed a convolutional neural network-gated recurrent unit hybrid framework for MTL. Wang et al. (2024) proposed a cross-modal hierarchical fusion approach for multimodal sentiment analysis, and Li et al. (2023) presented a graph-based self-supervised MTL method for EEG emotion recognition. Ling et al. (2024) presented a single-stage MTL attention network, and Shang et al. (2025) devised an interpretable spatio-temporal MTL framework. Despite these advances, applications of MTL in the petroleum sector remain unexplored. The tasks of system efficiency prediction and operating-condition classification for beam-pumping wells are inherently coupled; yet existing methods employ separate single-task models, resulting in redundant sensor-data processing, ignored inter-task physical correlations, and an inability to share diagnostic features—thereby weakening feature discriminability. Moreover, deploying two independent models doubles the computational burden on edge devices and constrains real-time responsiveness. Therefore, it is essential to develop a unified MTL model for simultaneous system efficiency prediction and condition diagnosis in beam-pumping well systems.

3. Baseline multi-task learning model for real-time system efficiency prediction and condition diagnosis

Abnormal operating conditions of beam pumping units directly lead to a decline in system efficiency, while sustained low efficiency often results in the occurrence or progression of faults. This bidirectional causal relationship makes the two tasks inherently coupled in terms of physical mechanisms. Currently, both the system efficiency prediction model and the operating condition diagnosis model for beam pumping units are implemented using single-task approaches. However, such designs fail to capture the inherent physical correlations between efficiency prediction and condition diagnosis, thereby neglecting valuable prior knowledge. Furthermore, practical deployment requires maintaining two independent models, which increases system complexity and operational costs. To overcome these challenges, we propose two multitask learning frameworks: (1) a real-time system efficiency and condition diagnosis model based on multimodal dual-path fusion, and (2) a real-time system efficiency and condition diagnosis model based on multimodal cross-scale fusion. By learning the of the prediction and diagnosis tasks and performing joint optimization with a multi-task loss function, the two tasks constrain and promote each other, thereby automatically capturing and exploiting the coupling relationship between them.

3.1. Multi-task learning model based on multimodal dual-path fusion for real-time system efficiency prediction and condition diagnosis

The proposed multimodal dual-path fusion multitask learning model for real-time system efficiency prediction and operating condition diagnosis consists of a numerical modality-guided

branch and an auxiliary fusion branch. In this design, numerical data are treated as the primary source of information, while string data, temporal sequences, and spatial sequences are incorporated as auxiliary sources to support feature fusion. The overall architecture includes a feature extraction module, a numerical modality guidance module, an auxiliary fusion module, a dual-path feature enhancement and output module, and a multitask joint optimization module, thereby enhancing both system efficiency prediction and operating condition classification. The structural schematic of the model is shown in Fig. 1, and the main steps are summarized as follows:

Step 1. Numerical data characterization: Numerical data are often high-dimensional, noisy, and redundant. Directly inputting such raw observations into the model can lead to the curse of dimensionality, reduced training efficiency, and increased sensitivity to noise and outliers, thereby impairing generalization performance. To address this challenge, a three-layer pyramidal convolutional neural network is constructed for the numerical input data $X_i^{\text{num}} (i = 1, \dots, N_{\text{num}})$, enabling multiscale feature extraction. The resulting features $H_{1,2,3}$ are then employed as inputs to both the numerical modality-guided fusion branch and the auxiliary fusion branch. The mathematical formulation of the numerical feature extraction process is given as follows:

$$H_{1,2,3} = \sigma(\text{Conv1D}(X_i^{\text{num}})) \quad (1)$$

where $\text{Conv1D}(\cdot)$ is a one-dimensional convolutional neural network. $\sigma(\cdot)$ is the activation function, and $H_{1,2,3}$ is the extracted features of the numerical data.

Step 2. Spatial sequence data feature extraction: The spatial sequence data of wellbore trajectories, consisting of inclination and dogleg severity, exhibit strong sequential dependencies and global structural patterns. The multi-head attention mechanism of the Transformer enables direct computation of correlations between any two positions in the sequence, while different attention heads capture distinct trajectory characteristics. This design allows the model to effectively learn both the sequential order and the structural patterns associated with depth variations. Therefore, we propose a Transformer-based feature extractor for spatial sequence modeling.

However, the global nature of the Transformer's multi-head self-attention mechanism reduces its sensitivity to high-frequency local features. In addition, global correlation propagation may amplify noise and spread it across the network, potentially destabilizing the modeling process. Although conventional convolutional modules can alleviate these issues, they often suffer from high computational overhead, feature mixing, local detail smoothing, and limited scalability.

To overcome these limitations, we integrate a separable convolution module with the Transformer to construct a separable-convolution-Transformer hybrid feature extractor. These two feature extractors are applied independently to the spatial sequences to produce feature representations, denoted as H_{sp1} and H_{sp2} , which are then used as inputs for the auxiliary fusion process. The mathematical formulation of the complete feature extraction procedure is given as follows:

$$H_{\text{sp1}} = \text{TransformerEncoder}(X_{\text{sp}}) \quad (2)$$

$$H_{\text{pointsp}} = \text{PointwiseConv}(\text{DepthwiseConv}(X_{\text{sp}}, W_{\text{dw}}), W_{\text{pw}}) \in \mathbb{R}^{B \times D \times H \times W} \quad (3)$$

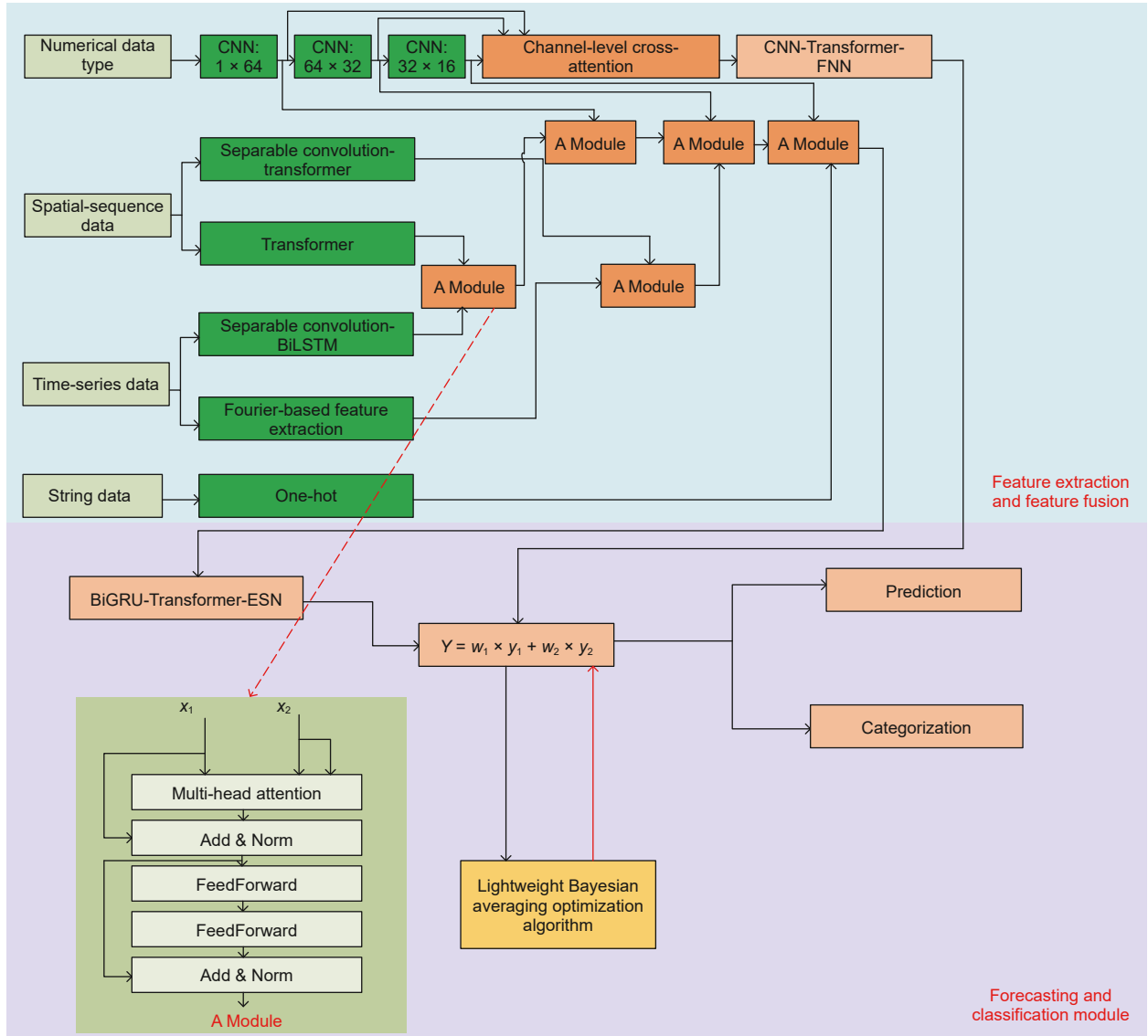


Fig. 1. Multi-task learning model for real-time system efficiency and working condition diagnosis based on multimodal two-path fusion.

$$H_{sp2} = \text{TransformerEncoder}(\text{Flatten}(H_{pointsp})) \quad (4)$$

where X_{sp} is the spatial sequence data. $\text{TransformerEncoder}(\cdot)$ is the Transformer mechanism. H_{sp1} and H_{sp2} is the spatial sequence features extracted by the two feature extractors. $\text{DepthwiseConv}(\cdot)$ is the depthwise convolution module. $\text{PointwiseConv}(\cdot)$ is the pointwise convolution. $\text{TransformerEncoder}(\cdot)$ is the Transformer module. $\text{Flatten}(\cdot)$ is the flattening module.

Step 3. Time series data feature extraction: Because time-series data exhibit prominent temporal patterns, along with frequency-domain characteristics and multi-scale dynamics, separable convolution is employed to efficiently extract short-term waveform features from the power sequences, while requiring considerably less computation than conventional convolutional neural networks. BiLSTM is particularly effective for modeling long-term trends and correlations across multiple stroke cycles. When combined in a cascaded architecture, these two modules enable a more comprehensive extraction of features from electrical power time-series data.

However, for certain abnormal power curves where waveform patterns are less distinguishable, the cascaded separable-convolution-BiLSTM feature extractor may fail to capture accurate representations. To address this limitation, a Fourier-based feature extraction method is introduced. These two feature extractors are applied independently to the time-series data, producing feature representations denoted as H_{T1} and H_{T2} , which are then used as inputs for the auxiliary fusion process. The mathematical formulation of the complete feature extraction procedure is expressed as follows:

$$H_{T1} = \frac{1}{T'} \sum_{t=1}^{T'} Z_{\text{BiLSTM}}[t] \in \mathbb{R}^{B \times 2H} \quad (5)$$

$$H_{T2} = |F(X)|W_{\text{fft}} + b \quad (6)$$

where Z_{BiLSTM} is the separable-convolution-BiLSTM feature extractor. H_{T1} and H_{T2} represent the extracted feature representations. $|F(X)|$ denotes the amplitude spectrum of the time series. b represents the bias term.

Step 4. String data feature extraction: One-hot encoding is applied to extract features from the string data, producing the feature representation H_s , which is subsequently used as an input to the auxiliary fusion process. The mathematical formulation of the string data feature extraction is given as follows:

$$H_s = \text{onehot}(X_s) \quad (7)$$

where $\text{onehot}(\cdot)$ is the encoder. H_s is the extracted features of the string data. X_s is the string data.

Step 5. Numerical modal data guided fusion: To enhance the discriminative capability of multi-scale features in numerical data, we propose an integrated mechanism that combines a channel-level interaction module with a multi-head attention mechanism. First, adaptive average pooling is applied to align the channel dimensions of the multi-scale feature maps, eliminating feature dimensional ambiguities caused by scale differences. The pooled features are then passed through the multi-head attention module for fusion, producing the fused representation B , which is subsequently used as the input for numerical-modal feature enhancement. The mathematical formulation of this fusion process is given as follows:

$$B = \text{softmax}\left(\frac{\text{AAP}(H_1)W_Q(\text{AAP}(H_2)W_K)^T}{\sqrt{d}}\right)\text{AAP}(H_3)W_V \quad (8)$$

where W_Q , W_K , and W_V is the trainable weight matrices. d is the scaling factor. $\text{AAP}(\cdot)$ is the adaptive average pooling module. H_j is the extracted features of the numerical data.

Step 6. Numerical modal data feature enhancement: To further enhance the interaction between local and global features, a CNN-Transformer-FNN cascaded feature enhancement layer is proposed. First, the CNN submodule is applied to extract local features from the output modeled by the multi-head attention module, followed by a flattening operation. Next, the Transformer establishes global dependencies and performs mean-pooling compression. Finally, a fully connected neural network (FNN) performs the final nonlinear mapping to generate the output representation y_1 , which is subsequently used as the input for the probability-weighted multi-task output fusion. The mathematical formulation of the entire feature enhancement layer is expressed as follows:

$$y_1 = \sigma\left(\text{MP}(\text{Transformer}(\text{Flatten}(AW_c + b_c)))W_f + b_f\right) \quad (9)$$

where W_c and b_c are the convolutional weights and biases. W_f and b_f are the fully connected layer weights and biases. y_1 is the feature output of the numerical modality. $\text{Flatten}(\cdot)$ is the flattening operation. $\text{MP}(\cdot)$ is the fully connected neural network. $\sigma(\cdot)$ is the activation function.

Step 7. Auxiliary fusion: Although the influencing features of system efficiency prediction and operating condition diagnosis for sucker-rod pumping wells originate from different sources and exhibit distinct scales, they collectively describe the of the pumping well and present strong with mutual interactions. Directly adding or concatenating these features may lead to. To mitigate this, the dual-input interactive attention fusion module (Module A) incorporates residual connections that add inputs directly to the outputs. This design preserves critical features during fusion and facilitates gradient propagation in deep networks. The combination of normalization and residual structures further alleviates issues such as gradient explosion or vanishing, ensuring stable convergence in the multimodal fusion process—even within deeply stacked architectures. Additionally, Dropout is applied to randomly mask a subset of feature units,

which reduces the model's dependence on any single modality or feature pattern and enhances generalization ability.

For these reasons, Module A is adopted as the core unit for feature fusion. First, the representations $\{H_{sp2}, H_{T1}\}$ and $\{H_{sp1}, H_{T2}\}$ are fused using Module A to produce the intermediate results Z_{output1} and Z_{output2} . Next, Z_{output1} and the first-layer numerical feature extraction result H_1 are progressively fused with Module A, yielding Z_{output3} . Then, Z_{output2} , H_2 , and Z_{output3} are further fused using Module A to generate a refined representation, Z_{output4} . Finally, Z_{output4} , the third-layer output H_3 of the multi-scale pyramid convolution block, and H_s are fused using Module A to produce the final representation Z_{output5} , which serves as the input for dual-path feature enhancement. The unified mathematical formulation of the fusion process in Module A is given as follows:

$$Z_{\text{ffn}}(H_{spi}, H_{spi}, H_{Tj}) = \text{FFN}\left(\text{LayerNorm}\left(Z + \text{Dropout}\left(\text{MultiHead}\left(H_{spi}, H_{spi}, H_{Tj}\right)\right)\right)\right) \quad (10)$$

$$Z_{\text{outputk}}(H_{spi}, H_{spi}, H_{Tj}) = \text{LayerNorm}\left(Z + \text{Dropout}\left(Z_{\text{ffn}}\left(H_{spi}, H_{spi}, H_{Tj}\right)\right)\right) \quad (11)$$

where $\text{Dropout}(\cdot)$ is the overfitting suppressor. $\text{MultiHead}(\cdot)$ is the multi-head attention mechanism. $\text{FFN}(\cdot)$ is the fully connected neural network layer. $\text{LayerNorm}(\cdot)$ is the batch normalization layer.

Step 8. Dual-path feature enhancement: Because the prediction and classification layers are directly connected to the feature fusion module, they rely exclusively on fused static features. This reliance limits their ability to adapt to variations in operating conditions or periodic patterns. To address this limitation, we design a cascaded feature enhancement layer that integrates BiGRU, Transformer, and ESN modules. Specifically, the BiGRU captures local temporal patterns, the Transformer models global dependencies across the sequence, and the ESN performs lightweight nonlinear mapping to generate the final representation y_2 . The mathematical formulation of the complete feature enhancement layer is presented as follows:

$$Z_{\text{Transformer}} = H_{\text{BiGRU}} + \text{MultiHead}(H_{\text{BiGRU}}) + \text{FNN}(H_{\text{BiGRU}} + \text{MultiHead}(H_{\text{BiGRU}})) \quad (12)$$

$$y_2 = W_{\text{out}}\left[Z_{\text{Transformer}}; \tanh\left(W_{\text{in}}Z_{\text{Transformer}} + W_{\text{res}}r_{t-1}\right)\right] + b_{\text{out}} \quad (13)$$

where $\tanh(\cdot)$ is the activation function. $\text{MultiHead}(\cdot)$ is the multi-head attention mechanism. $\text{FNN}(\cdot)$ is the fully connected neural network layer. H_{BiGRU} is the bidirectional GRU mechanism.

Step 9. Multi-task output fusion based on probability weighting: In the final output stage, the outputs from the dual-path feature enhancement module and the numerical-modal guidance module are integrated via lightweight Bayesian averaging optimization to achieve weighted fusion. The mathematical formulation of this process is provided below.

$$\omega_i \sim N(\mu_i, \sigma_i^2), i \in \{1, 2\} \quad (14)$$

$$y = \begin{cases} \text{Line}(w_1y_1 + w_2y_2), \text{ predicted values} \\ \text{softmax}(w_1y_1 + w_2y_2), \text{ categorical values} \end{cases} \quad (15)$$

where w_1 and w_2 are weights. $\text{Line}(\cdot)$ is the linear layer.

Step 10. Multi-task joint optimization: To mitigate overfitting and prevent weight over-concentration during Bayesian weighted optimization, a Kullback-Leibler (KL) divergence penalty is incorporated into the objective function. An adaptive loss weighting scheme is further employed to balance the optimization process. Moreover, to enhance both predictive and classification accuracy, three multi-task feature selection loss terms are designed and integrated into the overall loss formulation. The complete loss function is expressed as follows:

$$L_{\text{loss}} = 0.4L_{\text{Huber}} + 0.6L_{\text{class}} + \lambda_1 L_{L2} + \lambda_2 L_{\text{Fro}} + \lambda_3 L_{L1} + 0.1L_{\text{KL}} \quad (16)$$

where:

$$L_{\text{Huber}} = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \leq 0.8 \\ 0.8|y - \hat{y}| - \frac{1}{2}0.8^2, & |y - \hat{y}| > 0.8 \end{cases} \quad (17)$$

$$L_{L2} = \sum_{j=1}^d \sqrt{\sum_{i=1}^m W_{ij}^2} \quad (18)$$

$$L_{L1} = \sum_{i,j} |W_{ij}| \quad (19)$$

$$L_{\text{Fro}} = \sum_g \sqrt{\sum_{ij} (W_g)_{ij}^2} \quad (20)$$

$$\lambda_i = \lambda_{i,\text{max}} \cdot \left(1 - \frac{L_{\text{current}}}{L_{\text{initial}}}\right) \quad (21)$$

$$L_{\text{KL}} = \sum_{i=1}^2 \text{KL}\left(N(\mu_i, \sigma_i^2) \| N(0, \sigma_{\text{prior}}^2)\right) \quad (22)$$

where L_{L2} is L_2 regularization. L_{L1} is L_1 regularization. L_{Fro} is Frobenius penalty.

3.2. Multi-task learning model for real-time prediction of system efficiency and diagnosis of operating conditions based on multi-modal cross-scale fusion

For real-time efficiency prediction and condition diagnosis in beam-pumping well systems, electrical power time-series data serve as the primary source of dynamic information. However, a single modality alone is insufficient to effectively characterize complex operating conditions and nonlinear dynamic behaviors. To enhance feature representation capability, auxiliary information such as basic well parameters is incorporated to construct a multimodal feature set. Nevertheless, the significant differences in the distribution, scale, and structure of multimodal features can lead to challenges such as insufficient cross-modal correlation, dimensionality expansion, noise sensitivity, and scale inconsistency when directly concatenated for fusion, thereby constraining prediction accuracy and robustness. To address these issues, this study proposes a real-time multi-task learning model for efficiency prediction and condition diagnosis based on cross-scale multimodal fusion. The overall architecture, illustrated in Fig. 2, consists of four main components: a feature extraction module, a

feature fusion module, a feature enhancement and final output module, and a multi-task joint optimization module.

Step 1. Time series feature extraction: The power time series data exhibit complex characteristics such as pronounced local variations, significant long-term dependencies, non-stationarity, and multi-scale patterns. To address these challenges, this study employs BiTCN and BiLSTM as the fundamental feature extractors. Specifically, BiTCN is effective in capturing local dependencies within the sequence, whereas BiLSTM excels at modeling long-term temporal dependencies. To further integrate both local and global feature modeling, the time series is partitioned into three segments. Each segment is subsequently processed by BiTCN and BiLSTM feature extractors to capture complementary representations. The mathematical formulation of the time series feature extraction process is defined as follows:

$$\begin{cases} H_{T_j} = \text{BiTCN}(X_{T_i}), i = 1, 2, 3; j = 1, 2, 3 \\ H_{T_j} = \text{BiLSTM}(X_{T_i}), i = 1, 2, 3; j = 4, 5, 6 \end{cases} \quad (23)$$

where X_{T_1} , X_{T_2} , and X_{T_3} are the three segmented portions of the time-series data. H_{T_i} and H_{T_j} are the feature extraction results of segment i . $\text{BiTCN}(\cdot)$ is the bidirectional temporal convolutional network. $\text{BiLSTM}(\cdot)$ is the bidirectional long short-term memory network.

Step 2. Spatial sequence feature extraction: Spatial sequence data are typically represented either as continuous sequences of spatial coordinates or as spatial topological structures. Such data inherently exhibit local continuity while simultaneously embedding non-local spatial dependencies. To capture these characteristics, this study adopts CNN and BiLSTM as the fundamental feature extractors. Specifically, CNN, with their local receptive fields, are well-suited for identifying localized spatial patterns among adjacent positions, whereas BiLSTM are capable of modeling long-range dependencies across the spatial sequence. To further achieve a synergy between local and global feature representations, the spatial sequence is partitioned into three segments, with CNN and BiLSTM applied separately to each segment for feature extraction. The underlying mathematical formulation of the spatial sequence feature extraction process is presented as follows:

$$\begin{cases} H_{\text{sp}_j} = \text{CNN}(X_{\text{sp}_i}), i = 1, 2, 3; j = 1, 2, 3 \\ H_{\text{sp}_j} = \text{BiLSTM}(X_{\text{sp}_i}), i = 1, 2, 3; j = 4, 5, 6 \end{cases} \quad (24)$$

where H_{sp_j} and H_{sp_i} are the time-series feature extraction results. $\text{CNN}(\cdot)$ is the convolutional neural network. $\text{BiLSTM}(\cdot)$ is the bidirectional long short-term memory network.

Step 3. Numerical data feature extraction: Numerical data are often high-dimensional, noisy, and contain redundant raw observations. Directly feeding such data into the model can lead to the curse of dimensionality, reduced training efficiency, heightened sensitivity to noise and outliers, and diminished generalization performance. To mitigate these issues, a three-layer pyramid CNN is employed to perform multi-scale feature extraction on the numerical data, resulting in the extracted feature representation $H_{1,2,3}$. The mathematical formulation of this feature extraction process is given as follows:

$$H_{1,2,3} = \sigma \left(\text{Conv1D}_{\begin{matrix} 1 \rightarrow 64 \\ 64 \rightarrow 32 \\ 32 \rightarrow 16 \end{matrix}} (X_i^{\text{num}}) \right) \quad (25)$$

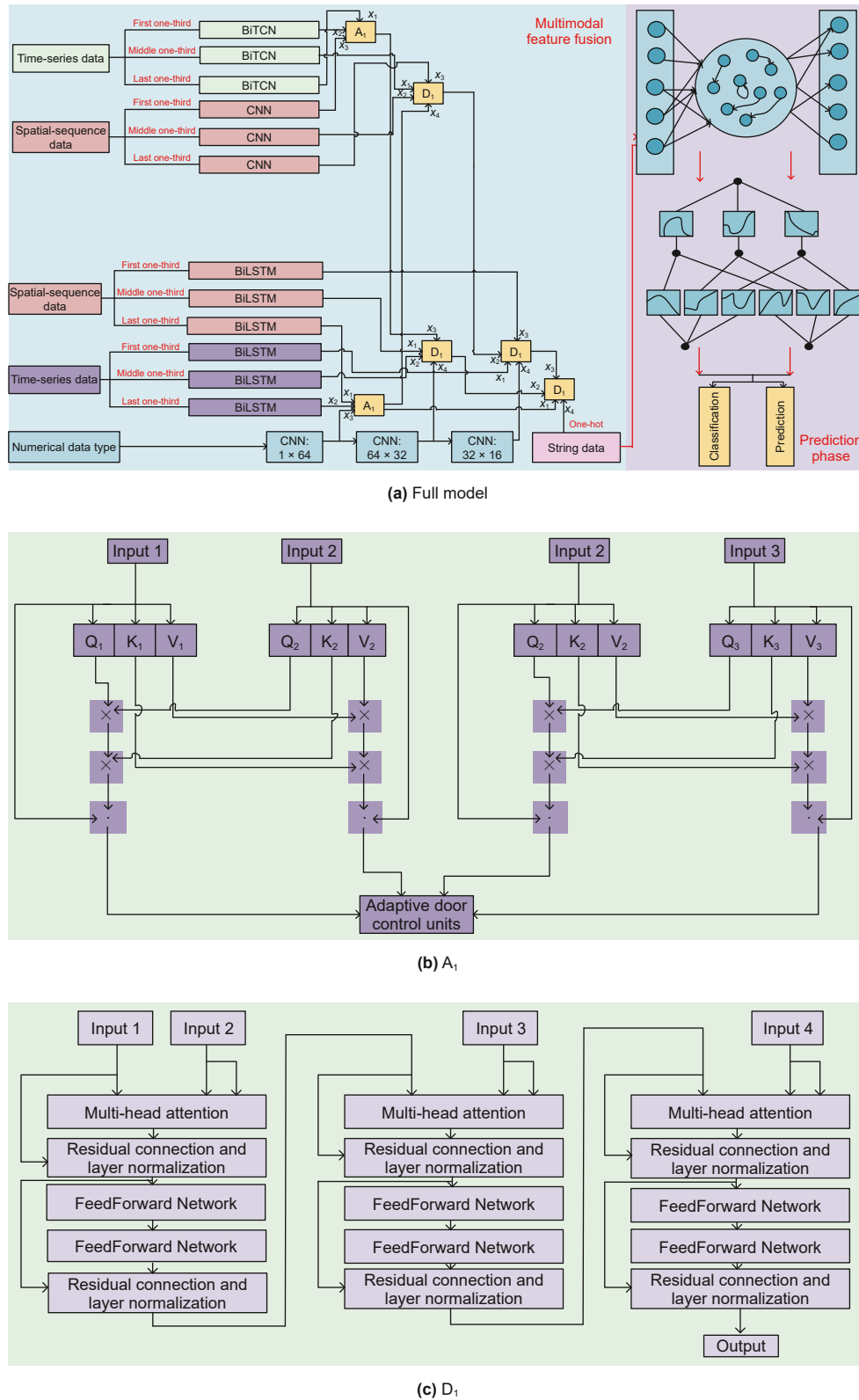


Fig. 2. Multi-task learning model for real-time prediction of system efficiency and diagnosis of operating conditions based on multi-modal.

where $\text{Conv1D}_{m \rightarrow n}(\cdot)$ is the one-dimensional convolutional neural network. $\sigma(\cdot)$ is the activation function. $H_{1,2,3}$ is to the extracted features of the numerical data. X_i^{num} is the numerical data.

Step 4. String data feature extraction: Since string data inherently lack the quantitative characteristics of numerical data, we employ one-hot encoding to extract their features, yielding the

feature representation H_s . The mathematical principle of the string data feature extraction is formulated as follows:

$$H_s = \text{onehot}(X_s) \quad (26)$$

where $\text{onehot}(\cdot)$ denotes the encoder, H_s represents the extracted features of the string data, and X_s denotes the string data itself.

Step 5. Initial feature fusion of upper and lower branches: In the feature fusion stage, to handle the complexity of temporal and spatial sequence data, a hierarchical interactive fusion model is proposed. In both the upper and lower branches, cross-branch scaled dot-product attention and adaptive gated fusion are utilized to integrate $\{X_{T_1}, X_{T_3}, X_{sp_1}\}$ and $\{H_1, H_{sp_6}, H_{T_6}\}$, producing the fused representations O_{out1} and O_{out2} . The overall fusion process can be expressed as follows:

$$H_{ij} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j, X_a, X_b, X_c \rightarrow (Q_i, K_i, V_i) \quad (27)$$

$$O_{out} = \sigma([H_b^{(a)}; \tilde{H}_b] W_g + b_g) \odot \tilde{H}_b + (1 - G_b) \odot H_b^{(a)} \quad (28)$$

where W_g is the weight, b_g is the bias. $\sigma(\cdot)$ is the activation function. $\odot$ is the element-wise product.

performance of both tasks within a unified representation space, a cascaded cross-attention block mechanism is employed to perform interactive fusion on $\{O_{out2}, Y_{L2}, Y_2, H_s\}$, yielding the final fused feature representation of all modalities, denoted as O_{out3} . The fusion process can be expressed as follows:

$$O_{out3} = \text{CAB}_3(\text{CAB}_2(\text{CAB}_1(O_{out2}, Y_{L2}), Y_2), H_s) \quad (32)$$

where $\text{CAB}_1(\cdot)$, $\text{CAB}_2(\cdot)$ and $\text{CAB}_3(\cdot)$ are the cascaded cross-attention block mechanism.

Step 9. Feature enhancement and final output: To enhance temporal modeling capability and nonlinear representation while maintaining computational efficiency, we propose a cascaded ESN-KAN integrated model for the global modeling of the cross-fused data. First, the ESN module captures the temporal dynamics of the input. Subsequently, the KAN network focuses on modeling global dependencies of s , thereby improving the accuracy of prediction and classification, with the final outputs denoted as y . The mathematical formulation of this global modeling process is given as follows:

$$s(t) = (1 - \rho)s(t-1) + \rho \tanh(W_{in}x(t) + W_{res}s(t-1)) \quad (33)$$

$$y = \begin{cases} \text{KAN}([s(1), s(2), \dots, s(L)]^T), \text{predicted values} \\ \frac{\exp(\text{KAN}([s(1), s(2), \dots, s(L)]^T) W_{cls} + b_{cls})}{\sum_{j=1}^C \exp(\text{KAN}([s(1), s(2), \dots, s(L)]^T) W_{cls} + b_{cls})}, \text{categorical values} \end{cases} \quad (34)$$

Step 6. Second feature fusion of upper and lower branches:

To facilitate complementary information exchange and improve model robustness, a cascaded cross-attention block mechanism is employed in both the upper and lower branches to enable interactive fusion between $\{O_{out2}, H_{T_2}, H_{sp2}, H_{sp3}\}$ and $\{O_{out2}, H_{T_4}, H_{sp5}, H_2\}$, yielding the fused outputs Y_{L1} and Y_{L2} . This fusion process can be uniformly formulated as:

$$Y_{L1} = \text{CAB}_3(\text{CAB}_2(\text{CAB}_1(O_{out2}, H_{sp2}), H_{sp3}), H_{T_2}) \quad (29)$$

$$Y_{L1} = \text{CAB}_3(\text{CAB}_2(\text{CAB}_1(O_{out2}, H_2), H_{sp5}), H_{T_4}) \quad (30)$$

where $\text{CAB}_1(\cdot)$, $\text{CAB}_2(\cdot)$ and $\text{CAB}_3(\cdot)$ are the cascaded cross-attention block mechanism.

Step 7. Lower layer branch scaling point multiplication self-attention feature fusion:

To adaptively capture critical dependencies within the feature sequence of the lower branch, thereby enhancing global intra-modal interactions, mitigating noise interference, and improving the discriminative capacity of feature representations, a cascaded cross-attention block mechanism is introduced in the lower branch. Through this mechanism, $\{H_3, H_{sp4}, H_{T_4}, Y_{L1}\}$ is interactively fused to generate the fused representation Y_2 . The fusion process can be expressed as:

$$Y_2 = \text{CAB}_3(\text{CAB}_2(\text{CAB}_1(H_3, H_{sp4}), H_{T_4}), Y_{L1}) \quad (31)$$

where $\text{CAB}_1(\cdot)$, $\text{CAB}_2(\cdot)$ and $\text{CAB}_3(\cdot)$ are the cascaded cross-attention block mechanism.

Step 8. Lower-level branch cascade cross-attention feature fusion:

To ensure that the model can simultaneously optimize the

where W_{in} and W_{res} are weights. C is the total number of categories. W_{cls} is the weight coefficient. b_{cls} is the bias.

Step 10. Multi-task joint optimization: To achieve a more balanced optimization, an adaptive weighted loss function is employed. In addition, to further enhance both prediction and classification accuracy, feature selection loss components designed for multi-task learning are integrated into the overall loss function. The complete formulation of the loss function is given as follows:

$$L_{loss} = \lambda_1 L_{RMSE} + \lambda_2 L_{class} + \lambda_3 L_{L2} \quad (35)$$

where L_{L2} is the L_2 -regularization term.

4. Multi-task learning model for real-time system efficiency and condition diagnosis based on meta-learning dynamic voting integration

Due to the limited generalization capacity, high variance sensitivity, insufficient robustness, poor interference resistance, limited adaptability to distribution drift, and constrained representational power of single multi-task learning models, we propose a meta-learning-based dynamic voting ensemble multi-task learning model for real-time system efficiency prediction and condition diagnosis, grounded in the balanced bootstrap theory. The schematic diagram is shown in Fig. 3. The procedural steps are as follows:

Step 1. First, the original dataset is divided into a training set X_{train} , a validation set X_{val} and a test set X_{test} .

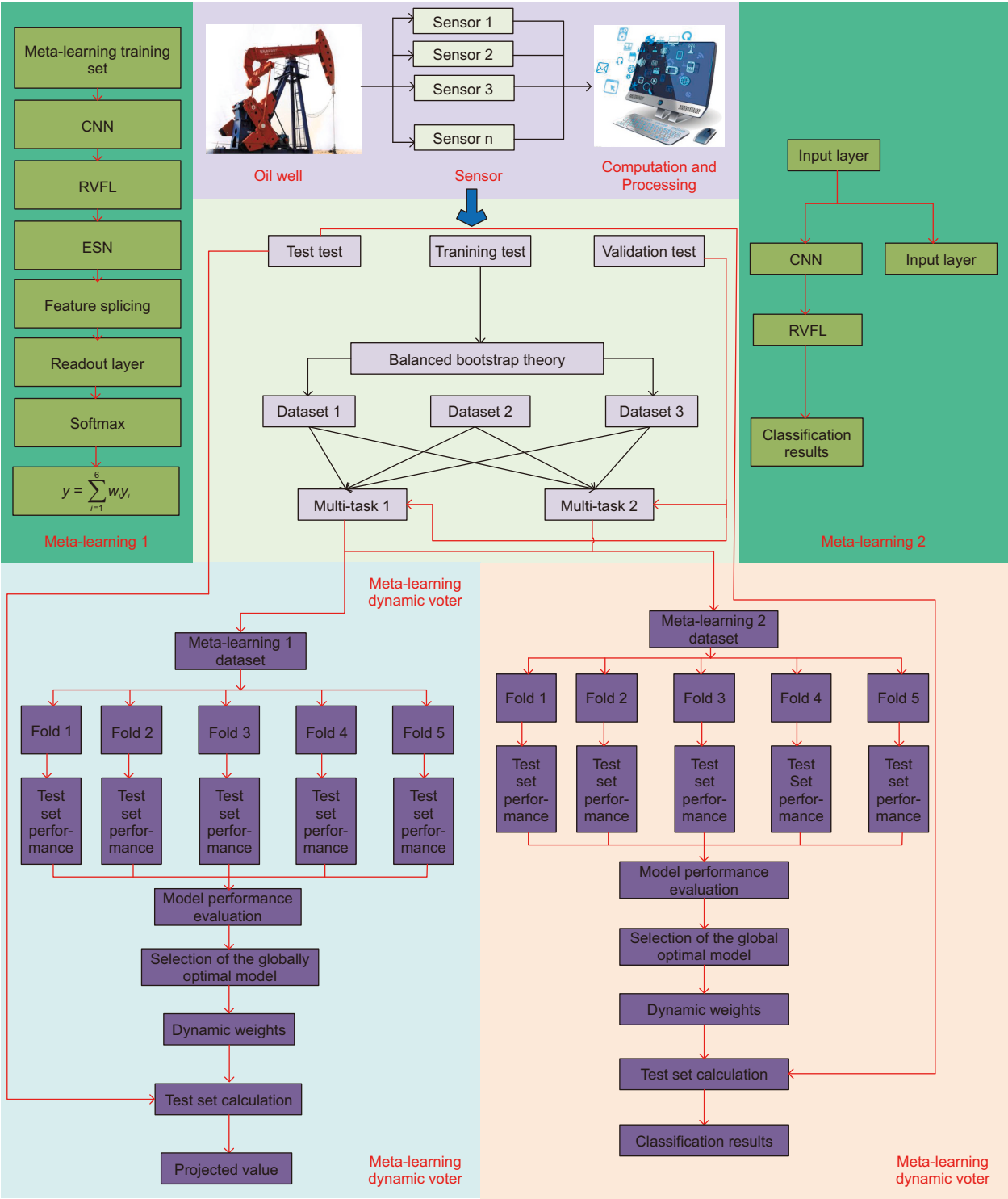


Fig. 3. Multi-task learning model for real-time system efficiency and condition diagnosis based on meta-learning dynamic voting integration.

Step 2. Then, to ensure the diversity of the underlying models, we extract three training subsets X_1 , X_2 and X_3 in the training set based on a balanced version of Bootstrap theory with putback.

Step 3. Second, the training subsets X_1 , X_2 , and X_3 are each used to train both the multimodal dual-path fusion and the multimodal cross-scale fusion multi-task learning models, yielding six base learners. These six base learners are then applied to the validation set X_{val} for prediction and classification, and their outputs are

concatenated to form the new meta-learner training sets X_{ytrain} and X_{ctrain} .

Step 4. To enhance the ensemble for efficiency prediction, we integrate five-fold cross-validation with a triple-feature learning mechanism and sample-adaptive dynamic base-model weighting. Specifically, the training set is partitioned into five folds. For each fold, a CNN-RVFL-ESN model is trained on its training subset and evaluated on the corresponding validation subset; the best-

performing model and its optimized weight coefficients from each fold are then aggregated to form the final ensemble for testing. The mathematical principle of the entire process is as follows:

$$M_j = \text{CNN} - \text{RVFL} - \text{ESN}(D_{\text{train}}^j) \quad (36)$$

$$\alpha_j = \arg\min_{\alpha} L_{\text{val}}^j(\alpha \cdot M_j(X_{\text{val}}^j)) \quad (37)$$

$$M_{\text{final}}^p = \text{Ensemble}(\{(\alpha_j, M_j)\}_{j=1}^5) \quad (38)$$

where $D_{\text{train}}^j, X_{\text{val}}^j$ denote the training set and validation set of the j -th fold, respectively; α_j represents the optimal weight coefficient of the model for the j -th fold; and M_{final}^p denotes the final ensemble model for efficiency prediction.

Step 5. Finally, for the condition-diagnosis ensemble, five-fold cross-validation is combined with a dynamic voting scheme. The training set is first partitioned into five folds. For each fold, a CNN-RVFL model is trained on its training subset and evaluated on the corresponding validation subset to obtain its validation accuracy. The model and weight coefficient that achieve the highest validation accuracy across all folds are then selected and used in a dynamic voting ensemble to perform final classification on the test set. The mathematical principle of the entire process is as follows:

$$M_j = \text{CNN} - \text{RVFL}(D_{\text{train}}^j) \quad (39)$$

$$\alpha_j = \arg\min_{\alpha} L_{\text{val}}^j(\alpha \cdot M_j(X_{\text{val}}^j)) \quad (40)$$

$$P_{\text{final}}^p = \text{Ensemble}(\{(\alpha_j, M_j)\}_{j=1}^5) \quad (41)$$

where $D_{\text{train}}^j, X_{\text{val}}^j$ denote the training set and validation set of the j -th fold, respectively; α_j represents the optimal weight coefficient of the model for the j -th fold; and P_{final}^p denotes the final ensemble model for probability prediction.

5. Experimental analysis

5.1. Data description

In this study, 3938 operational wells were randomly selected from an in-house oilfield database. The data comprises three types: numerical, string, and sequential. A summary of the sample features is provided in Table 1.

Analysis of Table 1 indicates that the dataset contains 31 features in total, including 30 influencing features and one target

feature. The data types are categorized as follows: “Pump Model” and “Balancing Method” are string variables. “Well Inclination”, “Dogleg Severity”, and “Electric Power Profile” are sequential variables. The remaining features are numerical. The sucker rod pumping system encompasses ten distinct operational conditions, namely: Upper Stuck Pump (USP:302), Lower Stuck Pump (LSP:304), Insufficient Supply (IS:297), Standing Valve Leakage (SVL:302), Rod Break (RB:317), Normal Operation (NO:995), Annular Leakage (AL:328), Gas Influence (GI:499), Gas Lock (GL:297), and Traveling Valve Leakage (TVL:297).

5.2. Data processing and evaluation indicators

To evaluate the predictive performance and diagnostic accuracy of the proposed multi-task learning model, the dataset was partitioned into training (70%), validation (15%), and test (15%) subsets. Given that each feature represents a distinct physical quantity, Z-score normalization was applied independently to the numerical, temporal-sequence, and spatial-sequence data, as defined below:

$$x = \frac{x - \mu}{\sigma} \quad (42)$$

where x is original feature. μ denotes the sample mean. σ denotes the sample standard deviation.

In order to be able to reasonably and multifacetedly evaluate the prediction and condition diagnosis accuracy of our proposed multi-task learning, we choose MSE, RMSE, R^2 and MAE as the evaluation metrics in the prediction problem. Their calculation formulas are shown below:

$$\text{MSE} = \frac{1}{M} \sum_{j=1}^M (y_{\text{ac},j} - y_{\text{pr},j})^2 \quad (43)$$

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{j=1}^M (y_{\text{ac},j} - y_{\text{pr},j})^2} \quad (44)$$

$$\text{MAE} = \frac{1}{M} \sum_{j=1}^M |y_{\text{ac},j} - y_{\text{pr},j}| \quad (45)$$

$$R^2 = 1 - \frac{\sum_{j=1}^M (y_{\text{ac},j} - y_{\text{pr},j})^2}{\sum_{j=1}^M (\bar{y}_{\text{ac},j} - y_{\text{ac},j})^2} \quad (46)$$

where M is the total number of samples. $y_{\text{ac},j}$ is the true values. $y_{\text{pr},j}$ is the predicted values.

Table 1

Data characterization.

Characteristics	Example	Characteristics	Example	Characteristics	Example
Pumping unit	CYJY10-3-48HB	Pumping speed	4 min ⁻¹	Downstroke	0.7 mm
Rated power	40 kW	Wellhead pressure	0.8 MPa	Rod break position	0 m
Saturation pressure	8 MPa	Casinghead pressure	0.6 MPa	Traveling valve leakage clearance	0 mm
Pumping unit	CYJY10-3-48HB	Pumping speed	4 min ⁻¹	Downstroke	0.7 mm
Density	800 kg/m ³	Water cut	15%	Standing valve leakage clearance	0 mm
Viscosity	65 mPa·s	Producing fluid level	600 m	Balancing method	Crank balance
GOR	100	Pump clearance	0	Number of centralizers	500
Specific gravity	0.6	Tubing diameter	62 mm		
Pump depth	1500 m	Sucker rod grade	3	Inclination angle	0.57, 0.57, 0.61, 0.61, 0.49, 0.54, ...
Pump diameter	38 mm	Equivalent diameter	23.06 mm	Dogleg severity	0, 6.75990, 5.9005, ...
Stroke length	3 m	Upstroke	0	Electric power	2.129928, 2.138211, 2.146494, ...
Efficiency	12.67%				

To evaluate the performance of operating condition diagnosis for beam pumping wells, this section introduces weighted average accuracy, weighted average precision, weighted average recall, and weighted average F1-score as evaluation metrics. The mathematical expressions of these metrics are as follows:

$$\text{Accuracy} = \frac{1}{N} \sum_{c=1}^{10} N_c \frac{TP_c + TN_c}{N_c} \quad (47)$$

$$\text{Precision} = \sum_{c=1}^{10} \left(\frac{N_c}{N} \times \frac{TP_c}{TP_c + FP_c} \right) \quad (48)$$

$$\text{Recall} = \sum_{c=1}^{10} \left(\frac{N_c}{N} \times \frac{TP_c}{TP_c + FN_c} \right) \quad (49)$$

$$F1 = \sum_{c=1}^{10} \left(\frac{N_c}{N} \times \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \right) \quad (50)$$

where TP_c , TN_c represent the true positives and true negatives for class c . Precision represents the precision for class c . N_c , N represent the total number of samples in class c and the overall sample size, Recall_c represents the recall for class c ; FP_c represents the false positives for class c ; FN_c represents the false negatives for class c .

5.3. Experimental analysis of multi-task learning based on multimodal dual-path fusion

To rigorously evaluate the effectiveness of the proposed multimodal dual-path fusion framework for pumpjack system efficiency prediction and working condition diagnosis within a multi-task learning paradigm, we established four baseline models for comparison: (i) a single-task system efficiency prediction model and (ii) a single-task working condition diagnosis model, both designed with the same shared-layer architecture as the proposed framework; (iii) a multi-task model based on attention mechanism fusion; and (iv) a multi-task model based on graph neural network fusion. To ensure the reliability of the experimental results, the dataset was split into 70% for training, 15% for validation, and 15% for testing. Both the proposed and baseline models were trained on the training set, with hyperparameters tuned on the validation set. The final evaluation was performed on the test set using the optimal hyperparameters identified during validation. All models were optimized using Adam with 100 training epochs, a batch size of 32, and a learning rate of 0.0001. In the spatial feature extraction module based on Transformer, the model feature dimension is set to 32, with 1 layer and 4 attention heads.

In the spatial feature extraction module combining separable convolution and Transformer, since the spatial sequence data include inclination angle and dogleg severity, the depthwise convolution of the separable convolution has an input channel number of 2, an output channel number of 2, a kernel size of 3×3 , a stride of 1, padding of 1, and uses ReLU as the activation function. The pointwise convolution has an input channel number of 2, an output channel number of 32, a kernel size of 1×1 , a stride of 1, and padding of 1. The Transformer model in this module has an input dimension of 32, 1 layer, and 4 attention heads.

In the temporal feature extraction module combining separable convolution and Transformer, since the time series data consist solely of electrical power time series, the depthwise convolution layer of the separable convolution has an input channel number of 1, an output channel number of 1, a kernel size of 3, a stride of 1, padding of 1, and uses ReLU as the activation function. The pointwise convolution has an input channel number of 1, an output channel number of 32, a kernel size of 1, a stride of 1, and a padding of 0, and also employs ReLU as the activation function. The BiLSTM module has an input dimension of 32, 1 layer, and a hidden dimension of 64.

In the cross-attention feature fusion module, the multi-head attention mechanism uses 4 heads, and the FeedForward neural network has a hidden dimension of 256, with both input and output dimensions set to 128. In the numerical data guidance module, the initial convolutional kernel size, with channel numbers of 64, 32, and 16 per layer. The channel attention mechanism uses 4 attention heads and an input channel number of 16. The subsequent convolutional layers consist of 2 layers, with a kernel size of 3×3 , a stride of 1, padding of 1, an input dimension of 16, and an output dimension of 128. The Transformer in this module has an input dimension of 128, 4 attention heads, and 1 layer.

In the BiGRU-Transformer-ESN architecture, the BiGRU component is configured with a single layer comprising 64 hidden neurons and accepts an input dimension of 128. The subsequent Transformer module operates with an input dimension of 128, a single layer, and 4 attention heads, utilizing ReLU as its activation function. The Echo State Network (ESN) is structured with 128 hidden neurons, a spectral radius of 0.95, a sparsity level of 0.1, and a leakage rate of 0.9, while processing inputs of dimension 1 and producing outputs of dimension 128. In the joint optimization loss function, the classification and regression losses were equally weighted at 0.5. To mitigate the effects of randomness, the experimental procedure was repeated five times, and the model performance was assessed based on the mean and standard deviation of these runs. As detailed in Table 2, this includes the evaluation metrics and inference times. Furthermore, Fig. 4 depicts the averaged loss function curves for each model. The averaged scatter plots representing system efficiency predictions and the confusion matrices for operational condition diagnosis are

Table 2
Mean and standard deviation of evaluation indicators.

Evaluation indicators	Multitasking model	Attention mechanism	Graph neural mechanism	Single-task prediction model	Single-task diagnostic model
R^2	0.8188 ± 0.0013	0.5180 ± 0.0096	0.4071 ± 0.0046	0.7244 ± 0.0141	
RMSE	5.8019 ± 0.0188	9.4636 ± 0.0948	10.4966 ± 0.0410	7.1547 ± 0.1835	
MSE	33.6771 ± 0.2371	89.5697 ± 1.7928	110.1799 ± 0.8622	51.2238 ± 2.615	
MAE	3.9764 ± 0.0149	6.7290 ± 0.0460	7.5616 ± 0.0247	4.9732 ± 0.1431	
Accuracy, %	94.22 ± 0.1200	76.11 ± 0.4100	70.93 ± 0.9200		92.19 ± 0.43
Precision, %	94.52 ± 0.1300	75.97 ± 0.4900	71.20 ± 1.1800		92.51 ± 0.40
Recall, %	94.22 ± 0.1200	76.11 ± 0.4100	70.93 ± 0.9200		92.19 ± 0.43
F1-score, %	94.26 ± 0.1200	75.27 ± 0.4500	70.18 ± 1.0700		92.27 ± 0.36
Time, ms	0.5514 ± 0.0215	0.5687 ± 0.01780	0.6085 ± 0.0399	0.5714 ± 0.0364	0.5732 ± 0.03
Memory, MB	6145.03	6109.95	6109.95	6145.02	6145.02

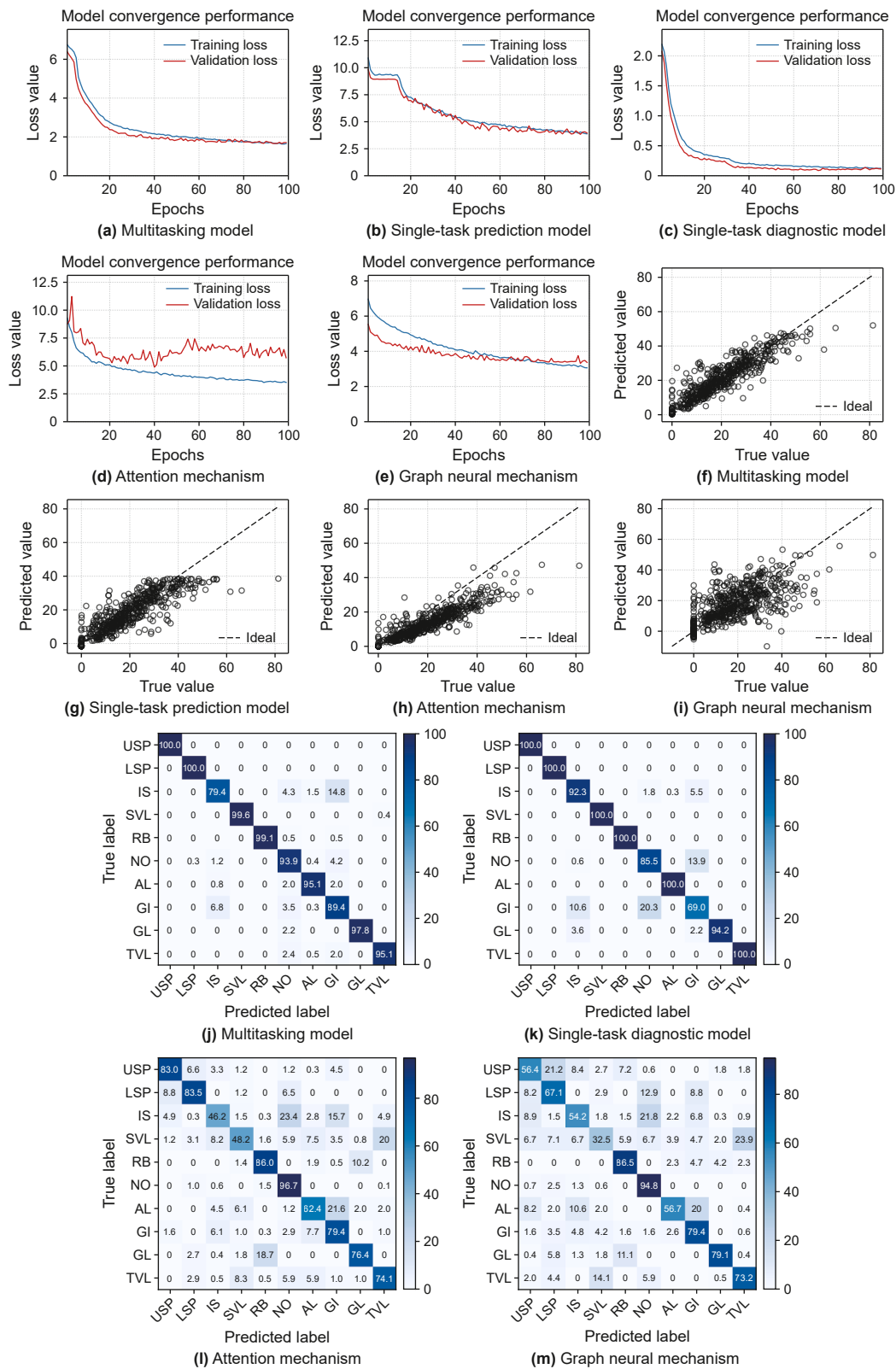


Fig. 4. Average scatter plot of system efficiency predictions and average operating condition diagnosis confusion matrix.

compiled in Fig. 4. Table 3 presents the percentage changes in evaluation indicators of the multitasking model compared to each of the comparison models.

As illustrated in Fig. 4, the average loss curves of all three models exhibit a monotonically decreasing trend on both the

training and validation sets, with the multi-task learning model demonstrating faster convergence and a lower stabilized loss. These observations indicate that all models were trained stably, without signs of significant overfitting or underfitting. Furthermore, compared to the single-task models, the multi-task learning

Table 3
Presents the percentage changes in evaluation indicators.

Evaluation indicators	Attention mechanism	Graph neural mechanism	Single-task prediction model	Single-task diagnostic model
R^2 , %	+58.07	+101.13	+13.03	
RMSE, %	−38.69	−44.73	−18.91	
MSE, %	−62.40	−69.44	−34.26	
MAE, %	−40.91	−47.41	−20.04	
Accuracy, %	+23.79	+32.84		+2.20
Precision, %	+24.42	+32.75		+2.17
Recall, %	+23.79	+32.84		+2.20
F1-score, %	+25.23	+34.31		+2.16
Time, %	−3.04	−9.38		−3.50

model showed accelerated convergence from the early stages of training and achieved a substantially lower final loss, suggesting that joint optimization effectively enhances the efficiency and generalization capability of feature learning.

Based on the findings presented in Tables 2 and 3, the multi-task model demonstrates substantial performance improvements across all evaluation metrics compared to the baseline models. Specifically, in terms of regression performance, the R^2 value exhibits a notable increase ranging from 13.03% to 101.13%, while the RMSE, MSE, and MAE are reduced by 18.91%–44.73%, 34.26%–69.44%, and 20.04%–47.42%, respectively. Regarding classification metrics, the Accuracy, Precision, Recall, and F1-score show enhancements between 2.20% and 32.84%, 2.17%–32.75%, 2.20%–32.84%, and 2.16%–34.31%, respectively. Furthermore, the inference time is also optimized, achieving a reduction of 3.04% to 9.38%.

5.4. Experimental analysis of multimodal cross-scale multi-task learning

To ensure the reliability of the experimental results, the dataset was divided into 70% for training, 15% for validation, and 15% for testing. Both the proposed model and the baseline models were trained on the training set, while hyperparameter tuning was conducted on the validation set. The optimal hyperparameter configuration identified through validation was then employed to evaluate model performance on the independent test set. During the experiments, we employed the Adam optimizer to train both the proposed and baseline models for 100 epochs with a batch size of 32 and a learning rate of 0.0001.

In the spatial feature extraction CNN module, designed to process the two spatial sequence features—inclination angle and dogleg severity—the CNN was configured with two input channels, a 3×3 kernel size, a hidden size of 64, padding of 1, adaptive average pooling, a fully-connected layer of 128 dimensions, and ReLU activation. The spatial BiLSTM module was constructed with an input size of 2, a single layer, and a hidden size of 64. For temporal feature extraction, the BiTCN module—processing only

the measured electrical power time series—was implemented with one input channel, 64 hidden channels, three stacked TCN blocks, a kernel size of 3, and a dilation factor of 1. The temporal BiLSTM module was correspondingly configured with one input channel, 64 hidden neurons, and a single layer.

Fusion module A1 was designed with a feature dimension of 128 and four gated inputs. Module D1 incorporated eight attention heads, with input, hidden, and normalization dimensions of 128, 512, and 128, respectively. The ESN module was parameterized with input, hidden, and output dimensions of 128, a spectral radius of 0.95, a sparsity of 0.1, and a leak rate of 0.9. In the joint optimization loss function, the weights for the classification and regression losses were set to 1 and 0.1, respectively.

To mitigate random variations, all experiments were repeated five times, with results reported as mean $\pm$ standard deviation. The evaluation metrics and inference times from these repetitions are summarized in Table 4. Fig. 5 displays the average loss curves for each model, while Fig. 5 presents the average scatter plots of system efficiency predictions and the average confusion matrices for working condition diagnosis. Table 5 presents the percentage changes in various evaluation metrics of the multi-task learning model compared to the baseline networks.

As shown in the average loss curves of different models in Fig. 5, both the multi-task and single-task models exhibited steadily decreasing training and validation losses as the number of iterations increased, eventually converging to stable values. These results indicate that the training processes of all models were stable, with no evident overfitting observed. Moreover, compared with the single-task models, the multi-task model achieved faster convergence and lower final loss values for both training and validation, demonstrating its superior feature learning and generalization capabilities.

Based on the results reported in Tables 4 and 5, the multi-task model achieves substantial improvements across all evaluation metrics compared to the baseline models. In terms of regression performance, the R^2 value shows a notable increase ranging from 3.66% to 38.61%, while the RMSE, MSE, and MAE are reduced by 6.86%–32.69%, 13.31%–54.78%, and 4.05%–45.16%, respectively. For

Table 4
Mean and standard deviation of evaluation indicators.

Evaluation indicators	Multitasking model	Single-task prediction model	Single-task diagnostic model	Attention mechanism
R^2	0.8131 $\pm$ 0.0021	0.7844 $\pm$ 0.0113		0.5866 $\pm$ 0.0383
RMSE	5.8937 $\pm$ 0.0334	6.3278 $\pm$ 0.1669		8.7553 $\pm$ 0.4120
MSE	34.7365 $\pm$ 0.3938	40.0683 $\pm$ 2.1015		76.8257 $\pm$ 7.1127
MAE	3.6705 $\pm$ 0.0518	3.8254 $\pm$ 0.1222		6.6933 $\pm$ 0.3890
Accuracy, %	97.97 $\pm$ 0.1500		95.77 $\pm$ 0.4400	35.84 $\pm$ 2.85
Precision, %	98.00 $\pm$ 0.1400		95.81 $\pm$ 0.4200	24.89 $\pm$ 6.80
Recall, %	97.97 $\pm$ 0.1500		95.77 $\pm$ 0.4400	35.84 $\pm$ 2.85
F1-score, %	97.96 $\pm$ 0.1500		95.75 $\pm$ 0.4400	24.97 $\pm$ 4.68
Time, ms	0.7744 $\pm$ 0.0158	0.7395 $\pm$ 0.0200	0.76248 $\pm$ 0.0254	1.252 $\pm$ 0.032
Peak memory usage, MB	10710.48	10710.48	10710.48	10621.78 $\pm$ 0.00

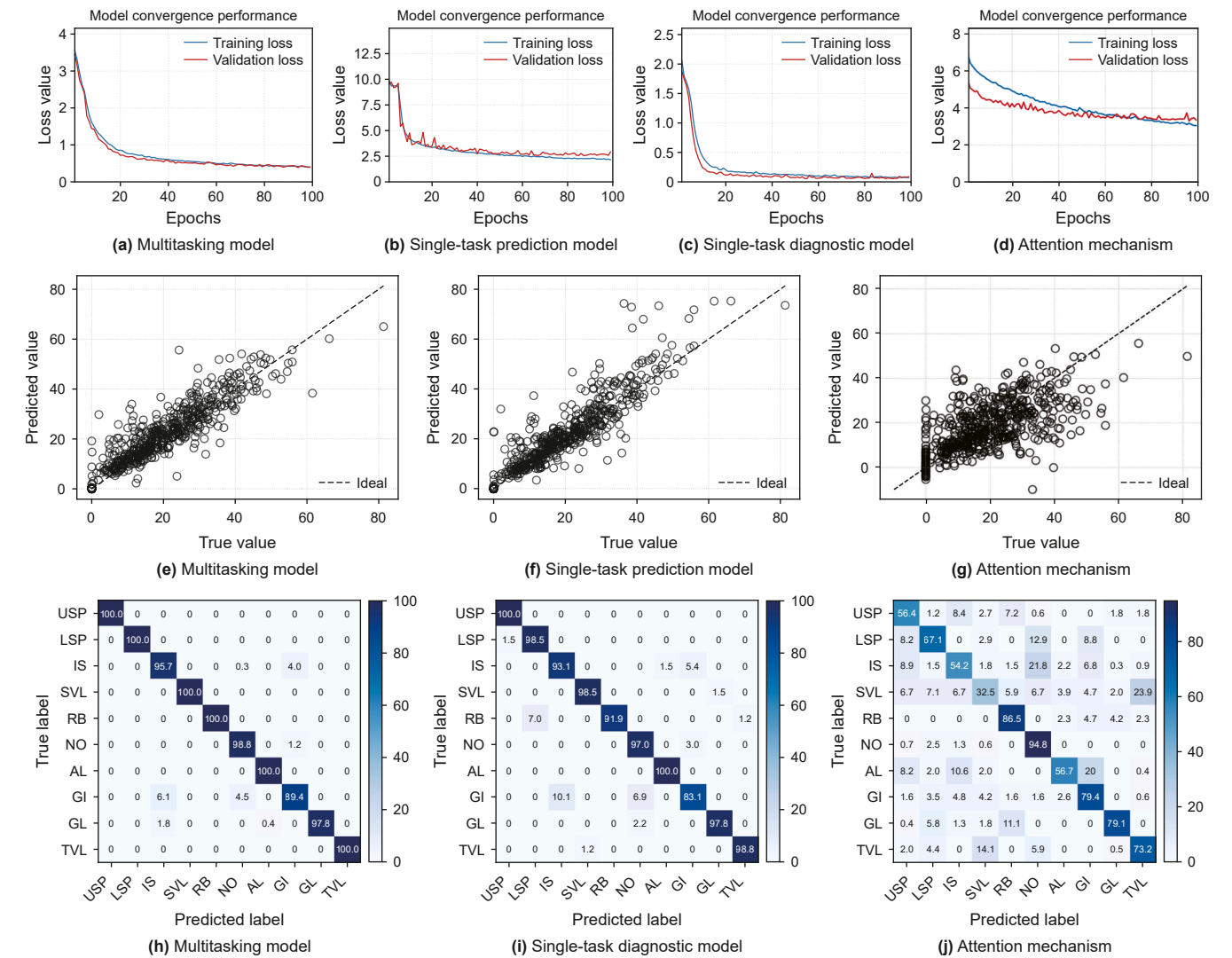


Fig. 5. Average scatter plot of system efficiency predictions and average operating condition diagnosis confusion matrix for model.

Table 5
Presents the percentage changes in evaluation indicators.

Evaluation indicators	Single-task prediction model	Single-task diagnostic model	Attention mechanism
R^2 , %	+3.66		+38.61
RMSE, %	−6.86		−32.69
MSE, %	−13.31		−54.78
MAE, %	−4.05		−45.16
Accuracy, %		+2.30	+173.35
Precision, %		+2.89	+293.73
Recall, %		+2.30	+173.35
F1-score, %		+2.31	+292.31
Time, %	+4.72	+1.56	+38.15

classification metrics, Accuracy, Precision, Recall, and F1-score are enhanced by 2.30%–173.35%, 2.89%–293.73%, 2.30%–173.35%, and 2.31%–292.31%, respectively. Furthermore, inference time is optimized, achieving a variation ranging from 1.56% to 38.15%.

5.5. Experimental analysis of multi-task learning based on meta-learning dynamic voting ensemble

To effectively evaluate the real-time system efficiency prediction and working condition diagnosis performance of the

proposed meta-learning-based dynamic voting ensemble method, we first adopted multimodal dual-path fusion multi-task learning and multimodal cross-scale fusion multi-task learning as fundamental components. Subsequently, based on the balanced bootstrap theory, three training subsets were generated, upon which a total of six baseline models were constructed for comparative analysis. To ensure reliable experimental results, 70% of the dataset was used for training, 15% for validation, and the remaining 15% for testing. Both the proposed model and the baseline models were trained on the training set, with hyperparameter tuning carried

out on the validation set. The optimal hyperparameter configuration obtained from the validation set was then applied to the independent test set for the final performance evaluation. To ensure the reliability of the experiments, the multimodal dual-path fusion multi-task learning module was configured with 200 training epochs, a learning rate of 0.0001, and a batch size of 32.

In the spatial feature extraction module based on Transformer, the feature dimension was set to 32, with one layer and four attention heads. In the spatial feature extraction module combining depthwise separable convolution and Transformer, since the spatial sequence data consisted of well deviation angle and dogleg severity, the depthwise convolution layer was configured with two input channels, two output channels, a kernel size of 3×3 , a stride of 1, a padding of 1, and ReLU activation. The pointwise convolution layer was designed with two input channels, 32 output channels, a kernel size of 1×1 , a stride of 1, and padding of 1. The subsequent Transformer employed an input dimension of 32, one layer, and four attention heads.

In the temporal feature extraction module based on depthwise separable convolution and Transformer, since the temporal data only included the electric power time series, the depthwise convolution layer was configured with one input channel, one output channel, a kernel size of 3, a stride of 1, padding of 1, and ReLU activation. The pointwise convolution layer had one input channel, 32 output channels, a kernel size of 1, a stride of 1, padding of 0, and ReLU activation. The BiLSTM module had an input dimension of 32, one layer, and a hidden size of 64.

In the cross-attention feature fusion module, the multi-head attention mechanism employs 4 heads, while the feedForward network features a hidden dimension of 256, with both input and output dimensions set to 128. Within the numerical data guidance module, the initial convolutional layer utilizes a kernel size of with channel dimensions of 64, 32, and 16 across successive layers. The channel attention mechanism is configured with 4 attention heads and processes input with 16 channels. Subsequent convolutional layers consist of 2 layers with 3×3 kernels, a stride of 1, and padding of 1, transforming the input dimension from 16 to 128. The Transformer component in this module operates with an input dimension of 128, 4 attention heads, and a single layer.

In the BiGRU-Transformer-ESN module, the BiGRU was configured with one layer, 64 hidden units, and an input dimension of 128. The Transformer had an input dimension of 128, one layer, four attention heads, and ReLU activation. The ESN was configured with 128 hidden units, a spectral radius of 0.95, sparsity of 0.1, a leaking rate of 0.9, an input dimension of 1, and an output dimension of 128. Finally, in the joint optimization loss function module, the weights of the classification and regression loss functions were both set to 0.5.

To ensure the reliability of the experiments, the multimodal cross-scale fusion multi-task learning module was configured with 100 training epochs, a batch size of 32, and a learning rate of

0.0001. In the spatial feature extraction CNN module, since the spatial sequence data consisted of well deviation angle and dogleg severity, the number of input channels was set to 2, with a kernel size of 3×3 , hidden layer size of 64, padding of 1, adaptive average pooling, a fully connected layer of 128 dimensions, and ReLU activation. In the spatial feature extraction BiLSTM module, the number of input channels was 2, with one layer and a hidden size of 64.

For the temporal feature extraction BiTCN, as the time-series data only included measured electric power, the number of input channels was 1, with 64 hidden channels, three stacked TCN blocks, a kernel size of 3, and a dilation factor of 1. In the temporal feature extraction BiLSTM, the input dimension was 1, with one layer and 64 hidden units. In the fusion module A1, the feature dimension was set to 128 with four gated inputs. In the D1 module, the number of attention heads was 8, the input dimension was 128, the hidden dimension was 512, and the normalization dimension was 128.

In the ESN module, the architecture was configured with an input dimension of 128, a hidden layer dimension of 128, and an output dimension of 128. The network parameters were set with a spectral radius of 0.95, a sparsity of 0.1, and a leak rate of 0.9. For the joint optimization loss function, the weights for classification loss and regression loss were set to 1.0 and 0.1, respectively. The mean values and standard deviations of evaluation metrics for both the baseline multi-task learning models and the meta-learning-based dynamic voting integration approach for real-time system efficiency prediction and working condition diagnosis are summarized in Tables 6 and 7, respectively. Table 8 presents the relative performance changes of each baseline model compared to the integrated learning model in terms of average metric values. Fig. 6 illustrates the average loss curves across all models, while Fig. 7 displays the corresponding average scatter plots for system efficiency prediction. The average confusion matrices for working condition diagnosis of each model are shown in Fig. 8.

As shown in the average loss curves of the models in Fig. 6, both the multimodal dual-path fusion model and the multimodal cross-scale fusion model exhibited steadily decreasing training and validation losses as the number of iterations increased, eventually converging to stable values. These results indicate that the training processes of all models were stable, with no obvious overfitting observed, and that the models demonstrated strong generalization capability.

As shown by the mean values and standard deviations of the evaluation metrics in Tables 6 and 8, the proposed meta-learning-based dynamic voting ensemble multi-task model achieved substantially better predictive performance than the baseline prediction models, with consistently lower variability across all metrics. Specifically, compared with the baseline models, the ensemble multi-task prediction model improved R^2 by 5.4%–20.0%,

Table 6
Mean and standard deviation of prediction evaluation metrics.

	R^2	RMSE	MSE	MAE	Time, ms
Model-1					
Training subset-1	0.7837 ± 0.0132	6.3379 ± 0.1952	40.2070 ± 2.4587	4.0198 ± 0.1242	0.6290 ± 0.0459
Training subset-2	0.7779 ± 0.0184	6.4198 ± 0.2680	41.2858 ± 3.4162	4.0556 ± 0.1166	0.6404 ± 0.0238
Training subset-3	0.7503 ± 0.0246	6.8037 ± 0.3350	46.4020 ± 4.5711	4.3189 ± 0.1808	0.5930 ± 0.0288
Model-2					
Training subset-1	0.8546 ± 0.0091	5.1952 ± 0.1624	27.0165 ± 1.6984	3.5409 ± 0.0904	0.7467 ± 0.0167
Training subset-2	0.8302 ± 0.0076	5.6166 ± 0.1228	31.5607 ± 1.4001	3.8484 ± 0.0950	0.7548 ± 0.0290
Training subset-3	0.8331 ± 0.0146	5.5638 ± 0.2456	31.0164 ± 2.7072	3.7694 ± 0.1522	0.7460 ± 0.0287
Ensemble	0.9006 ± 0.0011	4.2996 ± 0.0224	18.4871 ± 0.1933	2.8017 ± 0.0122	6.1096 ± 0.0105

Table 7
Mean and standard deviation of diagnostic evaluation indicators.

	Accuracy, %	Precision, %	Recall, %	F1-score, %	Time, ms
Model-1					
Training subset-1	96.14 ± 0.7217	96.24 ± 0.6935	96.14 ± 0.7217	96.18 ± 0.7062	0.6290 ± 0.0459
Training subset-2	93.74 ± 1.5170	93.58 ± 1.7357	93.74 ± 1.5170	93.52 ± 1.7085	0.6404 ± 0.0238
Training subset-3	93.81 ± 1.9198	93.64 ± 2.0881	93.81 ± 1.9198	93.63 ± 2.0852	0.5930 ± 0.0288
Model-2					
Training subset-1	97.56 ± 0.1734	97.60 ± 0.1822	97.56 ± 0.1734	97.58 ± 0.1744	0.7467 ± 0.0167
Training subset-2	97.46 ± 0.2404	97.50 ± 0.2107	97.46 ± 0.2404	97.46 ± 0.2282	0.7548 ± 0.0290
Training subset-3	97.03 ± 0.4332	97.07 ± 0.4035	97.02 ± 0.4332	97.02 ± 0.4314	0.7460 ± 0.0287
Ensemble	99.29 ± 0.0013	99.33 ± 0.0014	99.27 ± 0.0017	99.30 ± 0.0016	5.4096 ± 0.0105
Peak memory usage, MB	Training subset-1	Training subset-1	Training subset-1		
Model-1	6145.03	6145.03	6145.03		
Model-2	10710.48	10710.48	10710.48		
Ensemble (prediction/diagnostic)	10836.48/10743.48				

Table 8
Change in average evaluation indicators, %.

	R^2	MSE	MAE,	Accuracy	Precision	Recall	F1-score
Model-1							
Training subset-1	14.9	32.2	54.0	30.3	3.28	3.21	3.26
Training subset-2	15.8	33.0	55.2	30.9	5.92	6.15	5.89
Training subset-3	20.0	36.8	60.2	35.1	5.84	6.08	5.81
Model-2							
Training subset-1	5.4	17.3	31.6	20.9	1.77	1.77	1.75
Training subset-2	8.5	23.4	41.4	27.2	1.87	1.87	1.86
Training subset-3	8.1	22.7	40.4	25.7	2.33	2.33	2.32

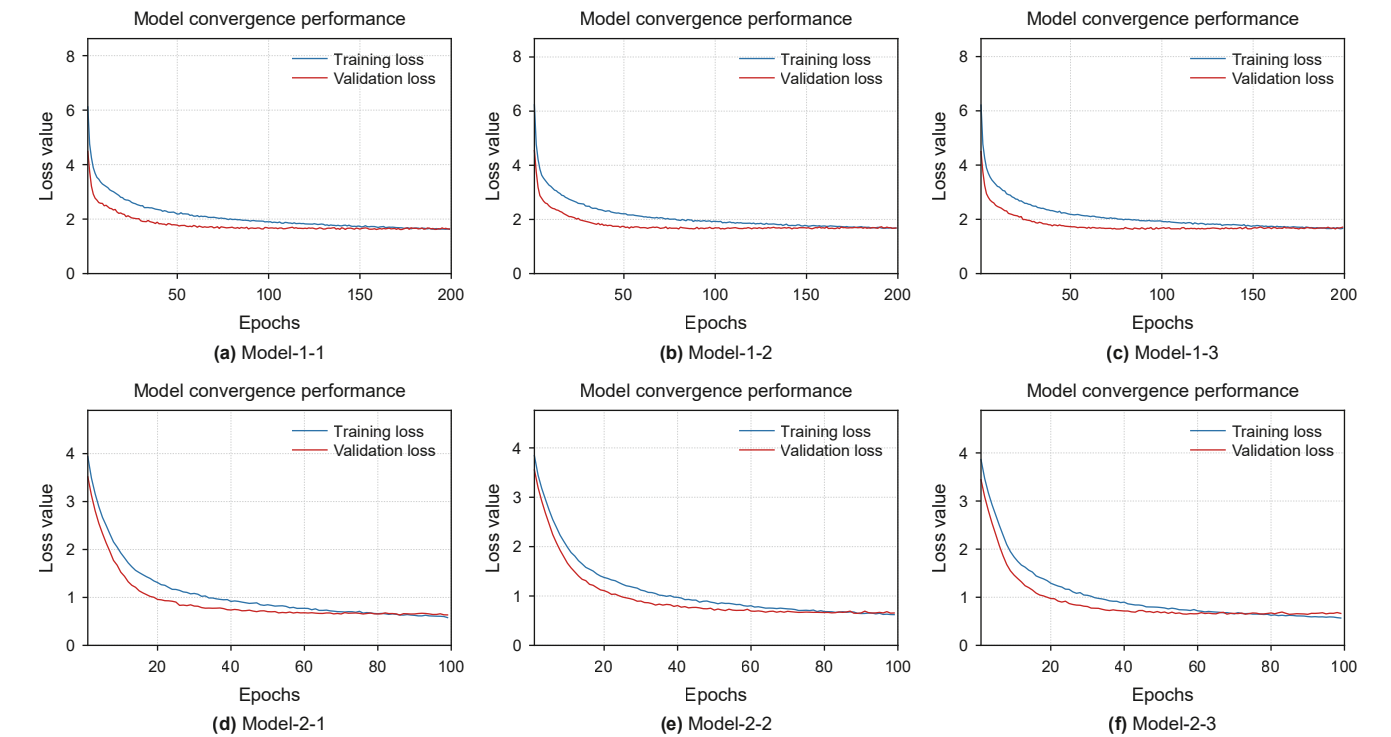


Fig. 6. Average loss functions of each base model.

reduced RMSE by 17.3%–36.8%, reduced MSE by 31.6%–60.2%, and reduced MAE by 20.9%–35.1%. The standard deviations of all evaluation metrics were also smaller for the ensemble model than for the baselines, indicating stronger stability. As illustrated in Fig. 7, the prediction scatter plots of the ensemble model were

more tightly clustered around the ideal diagonal, whereas the baseline models were more susceptible to the effects of working condition complexity and imbalanced sample distributions, leading to noticeable deviations in system efficiency prediction. These findings demonstrate that the proposed ensemble multi-task

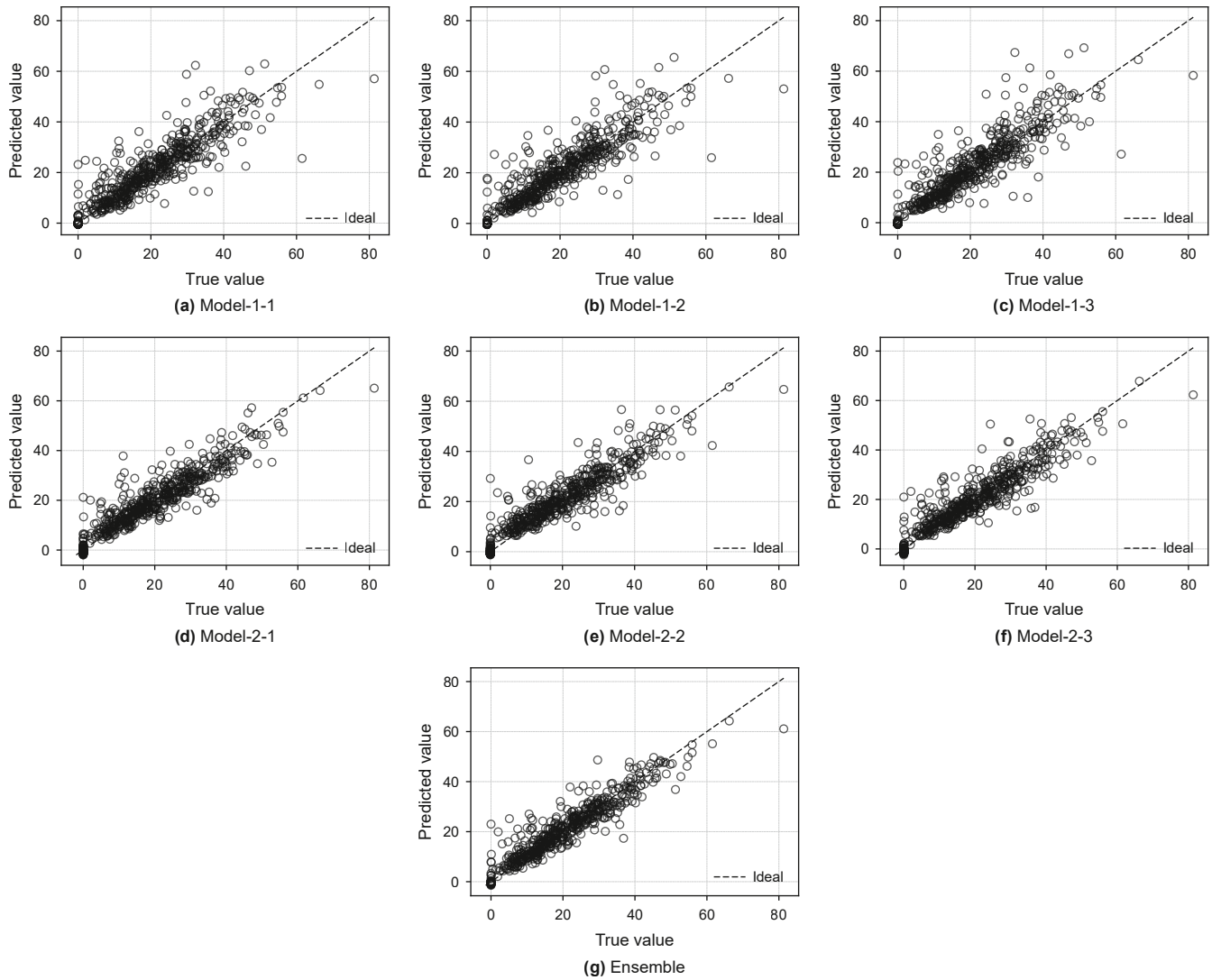


Fig. 7. Average prediction results for each model.

model outperforms the baseline models in overall predictive accuracy, generalization ability, and diagnostic reliability, thereby highlighting the practical value of ensemble learning for system efficiency prediction and working condition diagnosis in multi-task learning.

As shown by the mean values and standard deviations of the evaluation metrics in Tables 7 and 8, the proposed meta-learning-based dynamic voting ensemble multi-task model achieved markedly better diagnostic performance than the baseline diagnostic models, with consistently smaller standard deviations across all metrics. Specifically, compared with the baseline models, the ensemble multi-task model improved accuracy by 1.77%–5.92%, precision by 1.77%–6.15%, recall by 1.75%–5.89%, and F1-score by 1.76%–6.17%. The lower variability across these metrics further indicates that the proposed ensemble multi-task model is more effective in identifying different working conditions and demonstrates stronger robustness. As illustrated in Fig. 8, although the baseline models achieved relatively high recognition rates in most working condition categories, they still exhibited notable misclassifications in certain complex or boundary conditions, reflecting a degree of instability. In contrast, the ensemble model showed significant improvements in recognition performance across all categories, achieving nearly 100% accuracy overall.

Notably, it exhibited superior robustness and generalization ability in categories where the baseline models were less stable. These findings confirm that the proposed meta-learning-based dynamic voting ensemble method fully leverages the advantages of ensemble learning, thereby achieving more accurate and robust working condition diagnosis.

5.6. Multimodal dual-path fusion-based multi-task learning ablation experiment analysis

To systematically evaluate the effectiveness of individual components in the proposed multimodal dual-path fusion multi-task learning framework, we conducted a comprehensive ablation study. Specifically, in the spatial sequence feature extraction branch, the Transformer and depthwise separable convolution were removed in turn; in the temporal sequence branch, the BiLSTM and depthwise separable convolution were eliminated; in the CNN-Transformer module, the CNN component was excluded; and in the BiGRU-Transformer-ESN module, Transformer-ESN, BiGRU-ESN, Transformer, ESN, BiGRU, and Transformer were successively ablated. For the joint optimization loss function, we further assessed performance by removing L_1 , L_2 , and L_3 individually, as well as by excluding their pairwise combinations (L_1+L_2 ,

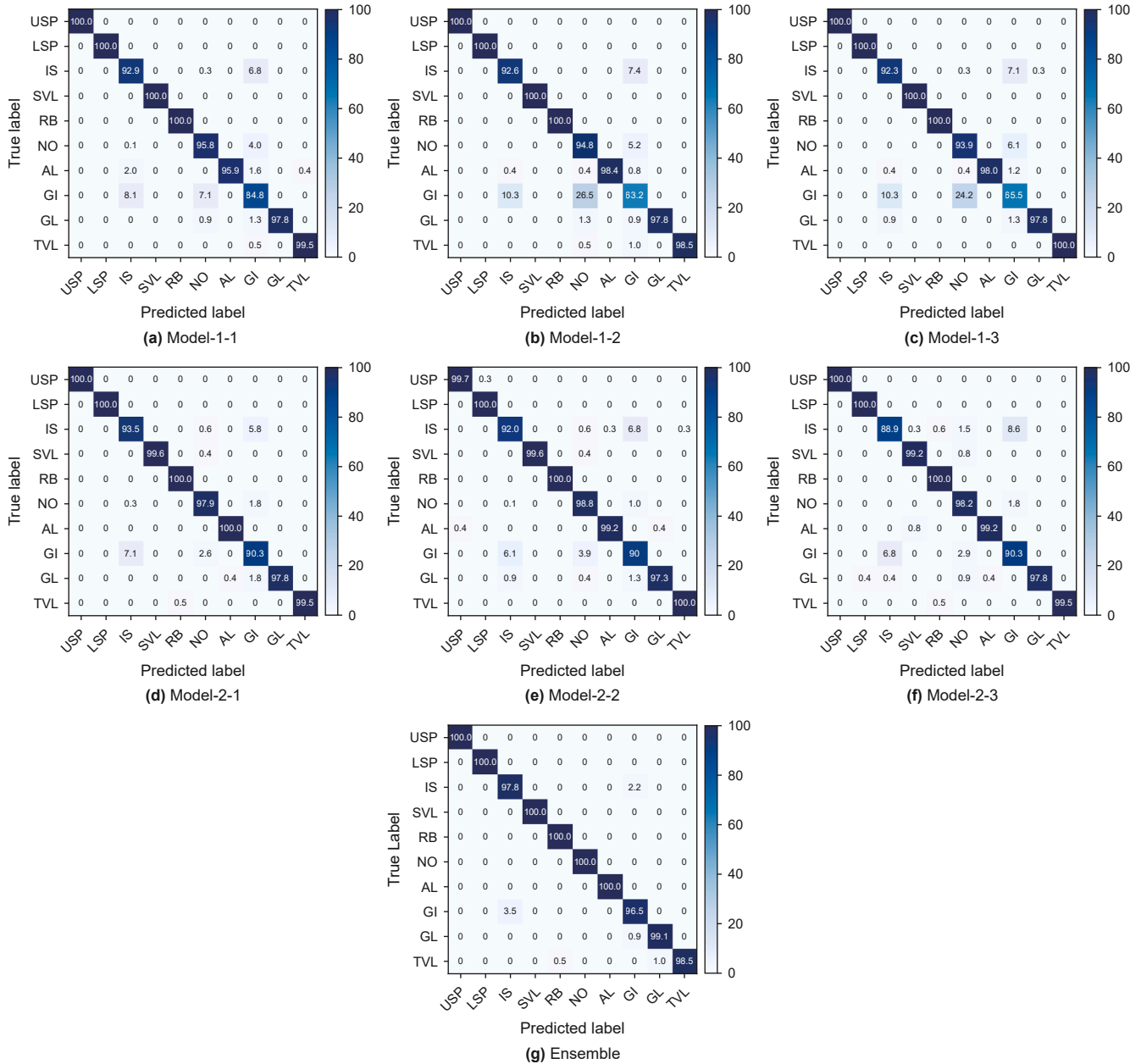


Fig. 8. Diagnostic results for each model.

L_1+L_3 , and L_2+L_3). To ensure the reliability of the experimental results, 70% of the dataset was allocated for training, 15% for validation, and the remaining 15% for testing. The proposed model and the baseline models were trained on the training set, while hyperparameter tuning was performed on the validation set. After obtaining the optimal hyperparameter configuration from validation, the test set was used to evaluate model performance. During training, the Adam optimizer was employed with 100 epochs, a batch size of 32, and a learning rate of 0.0001. The remaining structural hyperparameters followed those specified in Section 5.3. In the joint optimization loss function module, the weights of the classification and regression losses were both set to 0.5. To ensure statistical reliability and mitigate random effects, all experiments were conducted with five independent repetitions. The reported results represent the mean values and standard deviations derived from these replicates. As summarized in Tables 9 and 10, the evaluation metrics are presented together with their

corresponding standard deviations. Tables 11 and 12 present the percentage changes in each evaluation indicator.

As shown in Tables 11 and 12, compared with the ablation networks, the complete network achieves significant improvements across all evaluation metrics. Specifically, the MSE decreases by 11.77%–47.87%, RMSE decreases by 6.05%–27.78%, and MAE decreases by 3.77%–36.77%, while R^2 increases by 3.05%–25.51%. For the diagnostic performance, the Accuracy improves by 4.59%–53.73%, Precision improves by 16.52%–64.47%, Recall improves by 4.59%–53.73%, and F1-score improves by 7.76%–65.18%.

5.7. The impact of multi-task weighting coefficients on the model

Under the multi-task learning framework, loss functions from different tasks must be weighted and integrated for joint optimization. Inappropriate weight assignments can lead to one task

Table 9
Prediction evaluation metrics.

Model-1	MSE	RMSE	R^2	MAE
Multimodal dual-path fusion	33.6771 ± 0.2371	5.8019 ± 0.0188	0.8188 ± 0.0013	3.9764 ± 0.0149
Separable convolution-1	49.7054 ± 3.2804	7.0465 ± 0.2306	0.7326 ± 0.0177	4.6891 ± 0.1799
Transformer-1	42.0968 ± 1.8042	6.4867 ± 0.1400	0.7735 ± 0.0097	4.3861 ± 0.0514
Separable convolution-2	57.9755 ± 8.2852	7.5939 ± 0.5551	0.6881 ± 0.0446	5.0869 ± 0.3782
BiLSTM	46.5492 ± 1.5896	6.8217 ± 0.1166	0.7495 ± 0.0086	4.5536 ± 0.1007
Transformer-2	38.1678 ± 1.4346	6.1752 ± 0.1229	0.7946 ± 0.0077	4.1323 ± 0.0941
BiGRU	46.5117 ± 1.8700	6.8197 ± 0.1736	0.7497 ± 0.0247	4.5779 ± 0.1677
Transformer-3	55.7637 ± 2.3912	7.4658 ± 0.1612	0.7000 ± 0.0129	5.3447 ± 0.1203
ESN	64.6005 ± 3.9762	8.0336 ± 0.2482	0.6524 ± 0.0214	6.2893 ± 0.2677
BiGRU-Transformer	42.2224 ± 1.5604	6.4968 ± 0.1188	0.7728 ± 0.0084	4.5343 ± 0.0955
Transformer-ESN	43.9595 ± 1.4469	6.6293 ± 0.1095	0.7635 ± 0.0078	4.4667 ± 0.0847
BiGRU- ESN	41.3192 ± 1.5715	6.4269 ± 0.1227	0.7777 ± 0.0085	4.2039 ± 0.1116
L ₁	48.9831 ± 1.9721	6.9986 ± 0.1501	0.7364 ± 0.0106	4.8523 ± 0.1332
L ₂	47.1459 ± 1.1394	6.8658 ± 0.0831	0.7463 ± 0.0061	4.7736 ± 0.0570
L ₃	47.9204 ± 0.7898	6.9222 ± 0.0571	0.7422 ± 0.0042	4.8198 ± 0.0830
L ₁₋₂	43.1886 ± 0.6941	6.57008 ± 0.0495	0.7676 ± 0.0037	4.4000 ± 0.0822
L ₁₋₃	44.0295 ± 0.3110	6.6355 ± 0.0235	0.7631 ± 0.0017	4.4289 ± 0.0530
L ₂₋₃	41.3837 ± 0.5391	6.4328 ± 0.0475	0.7773 ± 0.0030	4.2854 ± 0.0180

Table 10
Diagnostic evaluation metrics.

Model-1	Accuracy	Precision	Recall	F1-score
Multimodal dual-path fusion	94.22 ± 0.1150	94.52 ± 0.1305	94.22 ± 0.1150	94.26 ± 0.1143
Separable convolution-1	61.29 ± 3.2193	57.47 ± 2.4342	61.29 ± 3.2193	57.07 ± 3.8100
Transformer-1	80.14 ± 1.4330	81.64 ± 1.2004	80.14 ± 1.4330	80.28 ± 1.3475
BiLSTM	64.10 ± 3.9878	66.96 ± 4.0931	64.10 ± 3.9878	61.50 ± 4.6252
Transformer-2	92.89 ± 0.4194	93.34 ± 0.4373	92.89 ± 0.4194	93.07 ± 0.4291
Separable convolution-2	69.48 ± 2.8747	72.52 ± 3.0996	69.48 ± 2.8747	67.97 ± 3.5644
BiGRU	87.31 ± 0.9093	87.84 ± 0.6688	87.31 ± 0.9093	87.48 ± 0.7969
Transformer-3	86.09 ± 1.7515	86.73 ± 1.4768	86.09 ± 1.7515	86.20 ± 1.7260
ESN	83.59 ± 1.1601	84.99 ± 1.0970	83.59 ± 1.1601	83.10 ± 0.9940
BiGRU-Transformer	88.97 ± 0.3877	89.20 ± 0.5190	88.97 ± 0.3877	89.03 ± 0.4129
Transformer-ESN	89.04 ± 0.6086	89.46 ± 0.2853	89.04 ± 0.6086	89.15 ± 0.4969
BiGRU- ESN	89.98 ± 0.8310	90.02 ± 0.7593	89.98 ± 0.8310	89.95 ± 0.8284
L ₁	87.65 ± 0.6387	87.62 ± 0.7625	87.65 ± 0.6387	87.36 ± 0.6687
L ₂	88.12 ± 0.6172	88.00 ± 0.7239	88.12 ± 0.6172	87.88 ± 0.6297
L ₃	87.38 ± 0.8872	87.53 ± 0.7464	87.38 ± 0.8872	87.24 ± 0.7932
L ₁₋₂	89.95 ± 0.3566	90.43 ± 0.5062	89.95 ± 0.3566	90.09 ± 0.4552
L ₁₋₃	90.02 ± 0.3877	90.81 ± 0.3304	90.02 ± 0.3877	90.13 ± 0.4736
L ₂₋₃	89.98 ± 0.4443	90.63 ± 0.4656	89.98 ± 0.4443	90.22 ± 0.0044

Table 11
Percentage change in prediction evaluation metrics.

Model-1	MSE, %	RMSE, %	R^2 , %	MAE, %
Separable convolution-1	−32.25	−17.66	+11.77	−15.20
Transformer-1	−20.00	−10.56	+5.86	−9.34
Separable convolution-2	−41.91	−23.60	+19.00	−21.83
BiLSTM	−27.65	−14.95	+9.25	−12.68
Transformer-2	−11.77	−6.05	+3.05	−3.77
BiGRU	−27.60	−14.93	+9.22	−13.14
Transformer-3	−39.61	−22.29	+16.97	−25.60
ESN	−47.87	−27.78	+25.51	−36.77
BiGRU-Transformer	−20.24	−10.70	+5.95	−12.30
Transformer-ESN	−23.39	−12.48	+7.24	−10.98
BiGRU- ESN	−18.50	−9.73	+5.29	−5.41
L ₁	−31.25	−17.10	+11.19	−18.05
L ₂	−28.57	−15.50	+9.72	−16.70
L ₃	−29.72	−16.19	+10.32	−17.50
L ₁₋₂	−22.02	−11.69	+6.67	−9.63
L ₁₋₃	−23.51	−12.56	+7.30	−10.22
L ₂₋₃	−18.62	−9.81	+5.34	−7.21

Table 12
Percentage change in diagnostic evaluation metrics.

Model-1	Accuracy, %	Precision, %	Recall, %	F1-score
Separable convolution-1	+53.73	+64.47	+53.73	+65.18
Transformer-1	+30.76	+42.06	+30.76	+40.67
Separable convolution-2	+13.37	+26.19	+13.37	+19.09
BiLSTM	+4.59	+16.52	+4.59	+7.76
Transformer-2	+51.57	+62.42	+51.57	+63.08
BiGRU	+42.45	+52.88	+42.45	+53.28
Transformer-3	+40.48	+50.94	+40.48	+51.07
ESN	+36.40	+47.89	+36.40	+45.60
BiGRU-Transformer	+45.18	+55.22	+45.18	+56.00
Transformer-ESN	+45.29	+55.67	+45.29	+56.21
BiGRU-ESN	+46.83	+56.65	+46.83	+57.63
L ₁	+43.01	+52.47	+43.01	+53.07
L ₂	+43.79	+53.13	+43.79	+53.97
L ₃	+42.56	+52.30	+42.56	+52.85
L ₁₋₂	+46.77	+57.35	+46.77	+57.86
L ₁₋₃	+46.90	+58.04	+46.90	+57.92
L ₂₋₃	+46.83	+57.71	+46.83	+58.08

dominating the training process, thereby hindering learning in other tasks. Thus, it is essential to analyze the weighting coefficients in multi-task learning to assess their impact on overall performance and inter-task balance. Specifically, for both the

multimodal dual-path fusion and multimodal cross-scale fusion multi-task learning models, two experimental setups were designed. In the first setup, the prediction task loss weight was fixed at 0.3, while the diagnosis task loss weight was varied at 0.1,

0.3, 0.6, and 0.9 to examine the effect of diagnosis task weighting on overall performance. In the second setup, the diagnosis task loss weight was fixed at 0.3, while the prediction task loss weight was adjusted to 0.1, 0.3, 0.6, and 0.9 to evaluate the influence of prediction task weighting.

To ensure experimental reliability, all trials were independently repeated five times, with the mean and standard deviation of each evaluation metric calculated from these repetitions. Fig. 9 illustrates how the mean and standard deviation of evaluation metrics vary with diagnostic and prediction loss weights in the

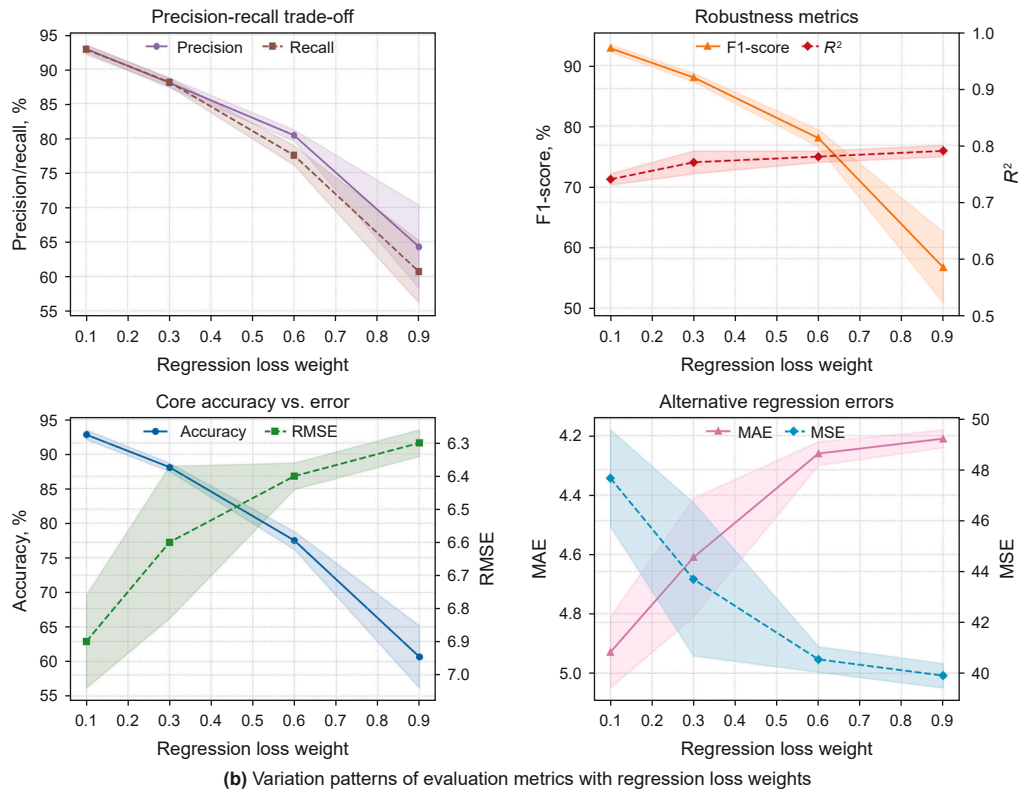
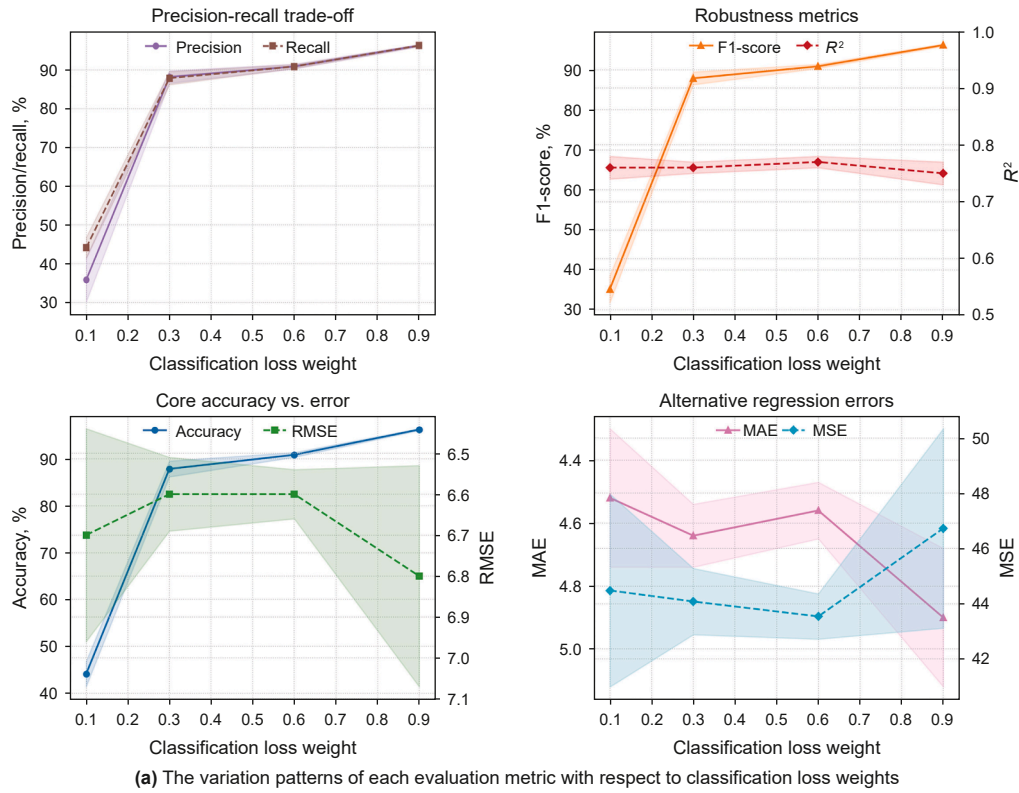
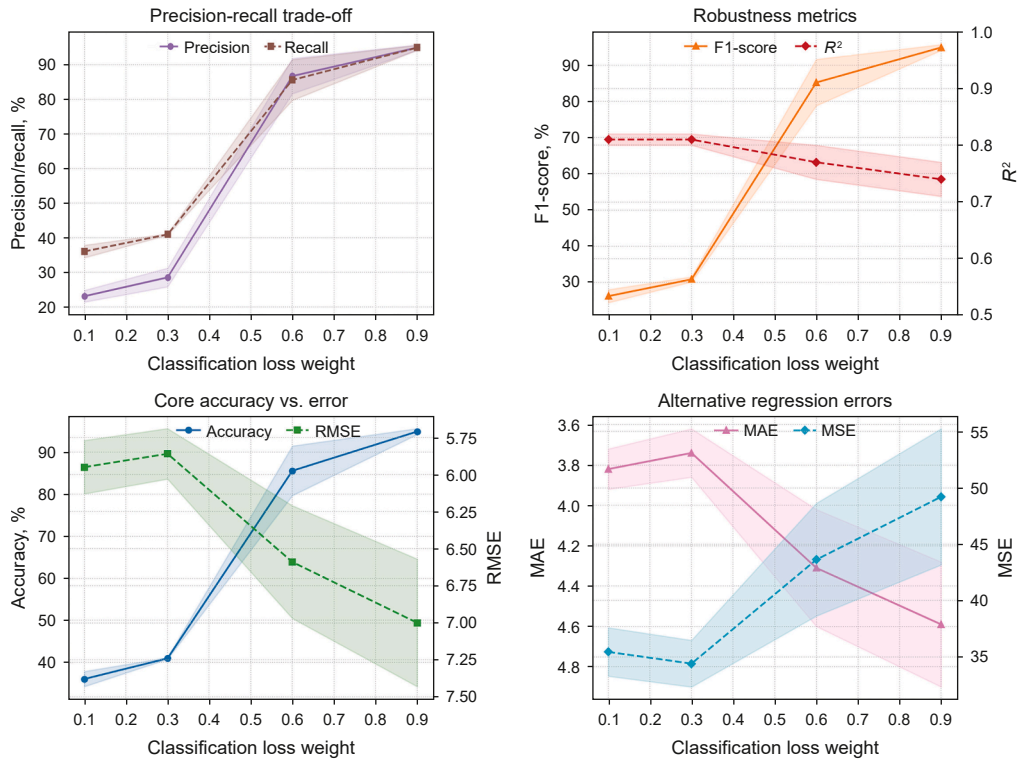


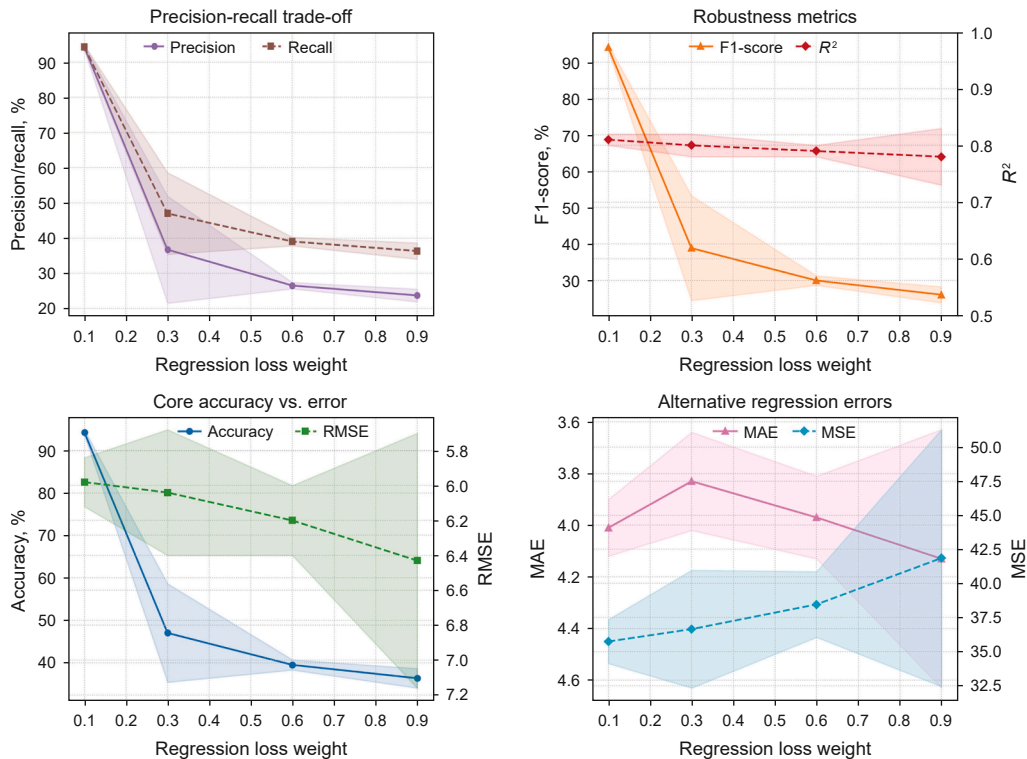
Fig. 9. Variation patterns of multi-modal dual-path fusion multi-task evaluation metrics with weight coefficients.

multimodal dual-path fusion model. Correspondingly, Fig. 10 presents the analogous variation trends observed in the multi-modal cross-scale fusion model under identical experimental conditions.

As observed from Fig. 9(a), with the increasing weight of the classification loss, the classification metrics generally exhibit a trend of initial improvement followed by stabilization, while the regression metrics progressively deteriorate. This pattern suggests



(a) The variation patterns of each evaluation metric with respect to classification loss weights



(b) Variation patterns of evaluation metrics with regression loss weights

Fig. 10. Variation patterns of multi-modal cross-scale fusion multi-task evaluation metrics with weight coefficients.

that when the classification weight is too low, the classification task lacks sufficient influence over the model; conversely, an excessively high classification weight tends to suppress the learning of the regression task. In contrast, Fig. 9(b) shows that as the regression loss weight increases, the regression metrics improve accordingly. However, this comes at the cost of a noticeable decline in classification performance, which drops sharply when the weight exceeds 0.6. These observations indicate a competitive relationship between the classification and regression tasks within the multi-task learning framework, underscoring the importance of appropriately balancing the weight coefficients to maintain both overall performance and inter-task equilibrium.

As shown in Fig. 10(a), increasing the classification loss weight leads to a clear upward trend in classification-related metrics, while the regression metrics exhibit fluctuations and even signs of deterioration. This suggests that a larger classification weight can substantially improve the model's representational capacity for classification, but excessive emphasis inevitably compromises its ability to capture regression patterns. Conversely, Fig. 10(b) shows that as the regression loss weight increases, regression metrics steadily improve, whereas classification performance declines markedly. These results demonstrate that larger regression weights enhance the model's fitting accuracy for regression tasks, but also highlight that in a multi-task framework, improper weight allocation may cause one task to dominate the optimization process at the expense of the others.

5.8. Discussion on engineering value

Due to the inherent difficulty in acquiring direct economic benefit data from oilfield operations, this section discusses the engineering value of the proposed methods based on existing literature and statistical data. The primary objective of online system efficiency prediction is to proactively identify wells with suboptimal efficiency, thereby enabling timely interventions to enhance operational performance. According to relevant statistics, a 5% improvement in system efficiency translates to average annual energy savings of 12,000 kWh per well. Consequently, accurate online prediction of system efficiency holds significant engineering value for energy conservation in beam pumping well operations. Furthermore, the accuracy of fault diagnosis in beam pumping wells substantially influences operation and maintenance decision-making. Statistical data from a Chinese oilfield indicate that a 1% reduction in the false alarm rate yields average annual energy savings of 5–10 kWh per well. Therefore, improving the diagnostic accuracy of beam pumping well conditions also contributes meaningfully to energy efficiency and holds considerable engineering value.

6. Conclusion

Since single-task models fail to capture the coupling relationship between efficiency prediction and condition diagnosis, this paper develops and compares three multi-task learning architectures for real-time analysis of beam pumping wells. First, the multimodal dual-path fusion model utilizes parallel branches to share features across heterogeneous modalities. Second, the cross-scale fusion model captures hierarchical interactions among features at different resolutions. Finally, the meta-learning based dynamic voting ensemble model adaptively integrates the outputs of sub-models to further improve prediction accuracy and diagnostic performance. Experimental results demonstrate that the proposed method effectively enhances the performance of both system efficiency prediction and working condition diagnosis models. Despite the remarkable performance achieved by the

proposed multi-task learning models, several limitations should be acknowledged. First, the high computational requirements of the current models may restrict their deployment on resource-constrained edge devices. Second, sensor reliability is of critical importance, as model performance degrades under conditions of sensor noise or signal dropout. Finally, the adaptability of the models to evolving field conditions has not yet been systematically evaluated. To address these issues, future research will focus on: developing lightweight architectures, exploring transfer learning across oilfields, and incorporating uncertainty quantification to enhance decision-making reliability under real-world uncertainties.

CRedit authorship contribution statement

Biao Ma: Investigation, Formal analysis, Conceptualization.
Shi-Min Dong: Writing – review & editing, Supervision, Investigation, Funding acquisition, Data curation, Conceptualization.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant: 51974276.

References

- Feng, Z.M., Wang, X., Huang, J.C., et al., 2025. Intelligent diagnostic method for beam pumping units based on torque indicator cards. *Measurement* 242, 115903. <https://doi.org/10.1016/j.measurement.2024.115903>.
- Gibbs, S.G., 1963. Predicting the behavior of sucker rod pumping systems. *J. Petrol. Technol.* 15 (7), 769–778. <https://doi.org/10.2118/588-PA>.
- Ju, C., Bibaut, A., Van, L., et al., 2018. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* 45, 2800–2818. <https://doi.org/10.1080/02664763.2018.1441383>.
- Lekia, S.D.L., Evans, R.D., 1965. A coupled rod and fluid dynamic model for predicting the behavior of sucker-rod pumping system. *SPE Prod. Facil.* 21664, 30–45. <https://doi.org/10.2118/30169-PA>.
- Li, J., Shao, J., Wang, W., et al., 2023. An evolutionary deep learning method based on multi-feature fusion for fault diagnosis in sucker rod pumping system. *Alex. Eng. J.* 66, 343–355. <https://doi.org/10.1016/j.aej.2022.11.028>.
- Li, Q., Chen, B., Huang, Z.Q., 2019. Study on equivalent viscous damping coefficient of sucker rod based on the principle of equal friction loss. *Math. Probl Eng.*, 9272751. <https://doi.org/10.1155/2019/9272751>.
- Ling, Y., Wang, Y., Dai, W., et al., 2024. MTANet: multi-Task attention network for automatic medical image segmentation and classification. *IEEE Trans. Med. Imag.* 43 (2), 674–685. <https://doi.org/10.1109/tmi.2023.3317088>.
- Lukasiewicz, S.A., 1991. Dynamic behavior of the sucker rod string in the Inclined well. *SPE production operations symposium*, 21665. <https://doi.org/10.2118/21665-MS>.
- Ma, B., Dong, S.M., 2023. Coupling simulation of longitudinal vibration of rod string and multi-phase pipe flow in wellbore and research on downhole energy efficiency. *Energies* 16 (13), 1–10. <https://doi.org/10.3390/en16134988>.
- Ma, B., Dong, S.M., 2024a. A hybrid prediction model for pumping well system efficiency based on stacking integration strategy. *Int. J. Energy Res.* 1–13, 8868949. <https://doi.org/10.1155/2024/8868949>.

- Ma, B., Dong, S.M., 2024b. A novel hybrid efficiency prediction model for pumping well system based on MDS-SSA-GNN. *Energy Sci. Eng.* 12 (8), 3272–3288. <https://doi.org/10.1002/ese3.1807>.
- Moreno, G.A., Garriz, A.E., 2020. Sucker rod string dynamics in deviated Wells. *J. Pet. Sci. Eng.* 184, 106534. <https://doi.org/10.1155/2019/3720127>.
- Shang, Y., Li, D., Li, Y., et al., 2025. Explainable spatiotemporal multi-task learning for electric vehicle charging demand prediction. *Appl. Energy* 384, 125460. <https://doi.org/10.1016/j.apenergy.2025.125460>.
- Sreenivasan, H., Krishna, S., 2024. Automatic sucker rod pump fault diagnostics by transfer learning using googlenet integrated machine learning classifiers. *Process Saf. Environ. Prot.* 191 (Part A), 14–26. <https://doi.org/10.1016/j.psep.2024.08.059>.
- Tan, C.D., Deng, H.W., Feng, Z.M., et al., 2022. Data-driven system efficiency prediction and production parameter optimization for PW-LHM. *J. Pet. Sci. Eng.* 209, 109810. <https://doi.org/10.1016/j.petrol.2021.109810>.
- Tarmigh, M., Behbahani, Hajidavalloo, E., et al., 2023. Two-way fluid–structure interaction for longitudinal vibration of a loaded elastic rod within a multi-phase fluid flow. *J. Braz. Soc. Mech. Sci. Eng.* 45, 572. <https://doi.org/10.1007/s40430-023-04484-4>.
- Wang, H.B., Dong, S.M., Zhang, Y., et al., 2023. Coupling simulation of the pressure in pump and the longitudinal vibration of sucker rod string based on gas-liquid separation. *Acta Pet. Sin.* 44 (2), 394–404. <https://doi.org/10.7623/syxb202302014> (in Chinese).
- Wang, H.B., Dong, S.M., 2020. Spatial transverse vibration simulation model of axially moving sucker rod string under the excitation of curved borehole. *Shock Vib.* 18, 3108718. <https://doi.org/10.1155/2020/3108718>.
- Wang, J., Feng, K., Wu, J., et al., 2019. SVM-based deep stacking networks. *AAAI Conference on Artificial Intelligence* 33, 5273–5280. <https://doi.org/10.1609/aaai.v33i01.33015273>.
- Wang, L., Peng, J., Zheng, C., et al., 2024. A cross modal hierarchical fusion multi-modal sentiment analysis method based on multi-task learning. *Inf. Process. Manag.* 61 (3), 103675. <https://doi.org/10.1016/j.ipm.2024.103675>.
- Wang, X.B., Lv, L.G., 2021. Longitudinal vibration analysis of sucker rod based on a simplified thermo-solid model. *J. Comput. Nonlinear Dynam.* 196, 107951. <https://doi.org/10.1016/j.petrol.2020.107951>.
- Wang, X., Liu, Y., Cheng, X., et al., 2025. ParaDC: Parallel-learning-based dynamometer cards augmentation with diffusion models in sucker rod pump systems. *Neurocomputing* 617, 128973. <https://doi.org/10.1016/j.neucom.2024.128973>.
- Wang, X., Kihara, D., Luo, J., et al., 2019. ENAET: Self-Trained Ensemble Autoencoding Transformations for Semi-supervised Learning. *ArXiv preprint, arXiv* 1911.09265. <https://doi.org/10.48550/arXiv.1911.09265>.
- Wei, J., Gao, X., 2020. Fault diagnosis of sucker rod pump based on deep-broad learning using motor data. *IEEE Access* 8, 222562–222571. <https://doi.org/10.1109/ACCESS.2020.3036078>.
- Xing, M.M., 2019. Response analysis of longitudinal vibration of sucker rod string considering rod buckling. *Adv. Eng. Software* 99, 49–58. <https://doi.org/10.1016/j.advengsoft.2016.05.004>.
- Xing, M.M., Zhou, L.L., 2015. Simulation analysis of nonlinear friction of rod string in sucker rod pumping system. *J. Comput. Nonlinear Dynam.* 14, 091008. <https://doi.org/10.1115/1.4044027>.
- Yang, X., Sun, J., Jin, B., et al., 2025. Multi-task aquatic toxicity prediction model based on multi-level features fusion. *J. Adv. Res.* 68, 477–489. <https://doi.org/10.1016/j.jare.2024.06.002>.
- Yin, J.J., Dong, S., Yang, Y.S., et al., 2021. Predicting multi-tapered sucker-rod pumping systems with the analytical solution. *J. Petrol. Sci. Eng.* 197, 108115. <https://doi.org/10.1016/j.petrol.2020.108115>.
- Zhang, C., Tu, L., Yang, Z., et al., 2025. A CMMOG-based lithium-battery SOH estimation method using multi-task learning framework. *J. Energy Storage* 107, 114884. <https://doi.org/10.1016/j.est.2024.114884>.
- Zheng, B.Y., Gao, X.W., 2017. Sucker rod pumping diagnosis using valve working position and parameter optimal continuous hidden markov model. *J. Process Control* 59, 1–12. <https://doi.org/10.1016/j.jprocont.2017.09.007>.