

Imbalanced Fault Diagnosis of Electrical Submersible Pumps Using a Graph Regularized Weighted Extreme Learning Machine

Wen Zhang, Li Deng

Southwest Petroleum University, Chengdu, Sichuan, 610500, China

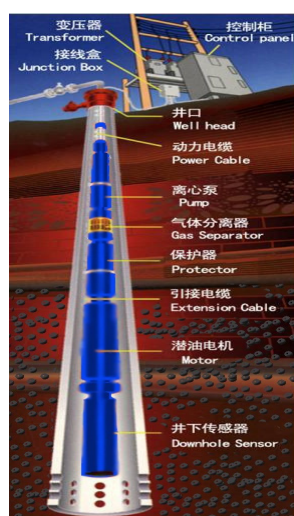
Abstract

To address class imbalance in spectrum-based fault diagnosis of electrical submersible pumps, a graph regularized weighted extreme learning machine is developed. Fixed features are obtained using random wavelet kernels and kernel group selection. The reciprocal of class frequency is used as the sample weight to balance the contributions of different classes to the supervised error. A graph Laplacian regularization term based on same-class neighbors is introduced to constrain the consistency of local predictions. The output weights are then obtained by solving a system of linear equations. Under a 32-fold protocol compatible with the dataset used, the method achieves an accuracy of 91.80%, a macro-averaged F1 score of 0.7500, and a balanced accuracy of 0.7810. Its macro-averaged F1 score exceeds that of the weighted extreme learning machine by 0.0496. Under controlled downsampling of the normal class and capped sampling of selected fault classes, the absolute improvement in macro-averaged F1 over WELM ranges from 0.0486 to 0.0575. The results indicate that class weighting and constraints on local structure within each class can jointly improve class-balanced diagnosis using fixed spectral features.

Keywords

Electrical Submersible Pump; Fault Diagnosis; Class Imbalance; Weighted Extreme Learning Machine; Graph Laplacian Regularization; Vibration Spectrum.

1. Introduction



(a) Schematic of the ESP system (b) Surface wellhead equipment and associated manifolds

Fig 1. ESP system structure and associated surface wellhead equipment

Electrical submersible pumps (ESPs) are important artificial lift equipment in oil and gas production, and their operating condition affects production continuity and equipment maintenance. Figure 1 shows the ESP system structure and associated surface wellhead

equipment. An ESP system comprises surface power supply and control equipment, together with downhole components such as the motor, protector, and pump. Vibration signals reflect changes in the condition of the rotor and shaft system. By mapping vibration features to fault labels, data-driven methods provide an effective means of identifying ESP faults [1].

In ESP fault diagnosis, the numbers of samples representing different operating conditions often vary. When normal samples dominate and some fault classes contain few samples, classification models tend to favor the majority class. High overall accuracy can therefore coexist with missed faults in minority classes. It is thus necessary to assess the recognition performance of individual classes alongside overall classification performance.

Existing approaches to learning from imbalanced data mainly adjust the sample distribution or modify the training objective to reduce the dominance of majority-class samples during training [2-4]. For example, SMOTE generates new samples by interpolating between minority-class samples to improve class proportions in the training data [5]. Meanwhile, ESP fault diagnosis has been investigated through feature learning and classification methods. Li et al. used a multivariate time-series tokenization Transformer to extract temporal information from process data [6], while Mello et al. improved fault classification through an ensemble of metric learners [7]. Related research has also noted that correlations among similar samples may cause conventional random splits to overestimate model performance, highlighting the need for an appropriate evaluation protocol [8]. These studies suggest that fault diagnosis requires discriminative features, consideration of class distributions during classifier training, and evaluation using suitable data partitions.

For feature extraction, random wavelet kernels can construct fault features from multiscale local signal responses [9]. However, obtaining features does not in itself resolve class imbalance: even with the same features, classifiers may differ in their recognition of majority and minority classes. This study therefore focuses on classifier design with fixed features. The extreme learning machine (ELM) trains a classifier through a random hidden-layer mapping and a solution for the output weights [10,11]. The weighted extreme learning machine (WELM) adjusts sample weights to balance the contributions of different classes to the training objective [12]. Class weighting, however, does not explicitly exploit local relationships among samples, particularly the structural information provided by samples with similar features within the same class. Graph regularization constrains prediction smoothness through sample adjacency and offers a way to exploit this information [13-15].

An unresolved issue is how to balance the training contributions of different classes while exploiting local relationships within each class to improve imbalanced fault diagnosis using existing spectral features. To address this issue, this study develops a graph regularized weighted extreme learning machine (LE-WELM) based on fixed features obtained through random wavelet kernels and kernel group selection. Inverse class-frequency weights balance the contributions of different classes, while a same-class neighbor graph encourages samples with similar features in the same class to have similar prediction outputs. By combining class weighting with local structural constraints, the study examines their complementary effects on minority fault recognition. Classifier comparisons, ablation studies, and controlled sampling experiments are used to evaluate the method.

2. Fault Diagnosis Using a Graph Regularized Weighted Extreme Learning Machine

LE-WELM is adopted to address the imbalanced distribution of ESP fault samples. Using fixed spectral features, the method applies class weighting to balance the contributions of different classes to the training objective and introduces graph regularization based on same-class neighbors to encourage consistent local predictions within each class. The aim is to improve

minority fault recognition. Output weights are obtained by solving a system of linear equations, and the supervised graph does not need to be reconstructed during prediction.

2.1. Fault Diagnosis Workflow

Figure 2 shows the LE-WELM fault diagnosis workflow. First, ESPset spectra are preprocessed according to the predefined folds, and fixed 1960-dimensional features are obtained using random wavelet kernels and GS-WKS. A fitting set is then constructed using either the natural distribution or a controlled sampling configuration. A random hidden-layer mapping produces the hidden-layer output matrix H , and the class-weight matrix W and same-class neighbor graph Laplacian L_g are computed. Model training is completed by solving a regularized system of linear equations for the output weights β . Finally, test features are fed into the model to predict fault classes, and diagnosis performance is evaluated on the fixed test set. Test labels are excluded from feature selection, parameter determination, class-weight calculation, and supervised graph construction throughout the process.

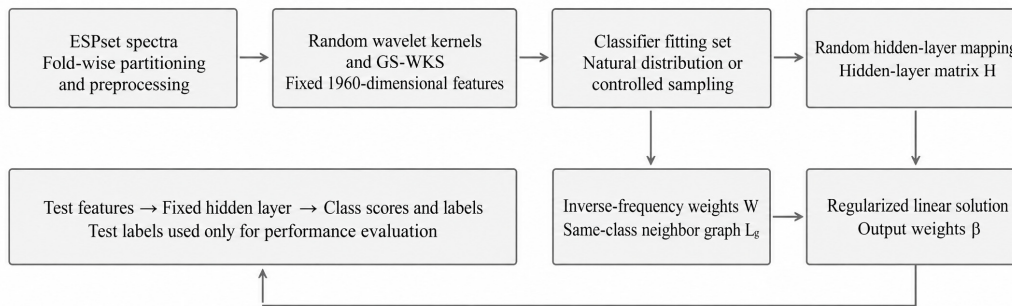


Fig 2. LE-WELM diagnosis workflow using fixed spectral features

2.2. Spectral Feature Representation

The original vibration spectra are high dimensional, and different frequency bands contain different fault information. Random wavelet kernels are therefore used to extract multiscale local features, and kernel group selection retains information that helps distinguish operating conditions, providing a common feature input for subsequent classification.

For the normalized vibration spectra in the dataset, the first 100 spectral points are removed, and the next 6100 consecutive points are used as input. The original and differenced spectra are processed using random wavelet kernels, multiple pooling statistics, and a secondary activation to form 2000 candidate features. The wavelet kernel group selection method (GS-WKS) selects features by importance separately within the two branches. A retention ratio of 0.98 yields a 1960-dimensional representation.

The feature mapping is fixed within each fold, and test data are excluded from parameter estimation and kernel group selection. All downstream classifiers use the same input features. Only the ablation without GS-WKS uses the unselected 2000-dimensional candidate features.

2.3. Random Hidden-layer Mapping and Class Weighting

To build a classifier that can be trained efficiently on the fixed features, a random hidden-layer mapping transforms the inputs into a nonlinear representation. Class weighting is also introduced to balance the training contributions of different classes, since majority-class samples can otherwise dominate the training loss.

Let the input matrix be $X \in \mathbb{R}^{N \times d}$, the number of hidden nodes be L , the random input weights be $A \in \mathbb{R}^{d \times L}$, and the hidden-layer biases be $b \in \mathbb{R}^L$. Following the random hidden-layer mapping of ELM [10,11], the hidden-layer output matrix is given by:

$$H = g(XA + 1_N b^T), \quad g(z) = \ln(1 + \exp(z)) \quad (1)$$

Here, $g(\cdot)$ is the Softplus activation used in this study, and the all-ones column vector has length N . A and b are sampled from uniform distributions over $[-1, 1]$ and $[0, 1]$, respectively, and remain fixed during output-layer training. The label matrix $T \in \mathbb{R}^{N \times 5}$ uses bipolar encoding: the entry corresponding to the true class is 1, and all other entries are -1 . With output weights $\beta \in \mathbb{R}^{L \times 5}$, the sample scores are $H\beta$.

For the i -th fitting sample, the inverse class-frequency weighting principle of WELM [12] assigns the reciprocal of its class frequency as its weight. The diagonal matrix W is constructed as follows:

$$w_i = \frac{1}{n(y_i)}, \quad W = \text{diag}(w_1, \dots, w_N) \quad (2)$$

Here, $n(y_i)$ is the number of samples in the fitting set whose label is y_i . No further global normalization is applied, so the sample weights sum to 1 within each class. The regularized weighted least-squares objective of WELM can be written as [12]:

$$\min_{\beta} \frac{1}{2} \|\beta\|_F^2 + \frac{C_1}{2} \|W^{\frac{1}{2}}(H\beta - T)\|_F^2 \quad (3)$$

Here, $C_1 > 0$ is the error-term coefficient, and $\|\cdot\|_F$ denotes the Frobenius norm. Class weighting reduces the influence of differences in class size on the loss, but does not yet exploit local geometric relationships among samples within the same class.

2.4. Construction of the Same-class Neighbor Graph

Class weighting adjusts each class's contribution to training but does not explicitly exploit sample similarity. A same-class neighbor graph is therefore constructed to connect samples with similar features within each class, providing a basis for constraining the consistency of their predictions.

For each fitting sample x_i , only the k nearest neighbors in the same class are considered. The local scale σ_i is the Euclidean distance to the t -th nearest same-class neighbor. If there are insufficient samples, k and t are capped at the available number of non-self samples. A zero scale is replaced by the median positive distance within the class; if no positive distance is available, a positive fallback value is used. The local-scale selection follows the self-tuning similarity principle [16], with connections further restricted to samples of the same class. The neighbor weights are:

$$Z_{ij} = \exp\left[-\frac{\|x_i - x_j\|_2^2}{2\sigma_i\sigma_j}\right] \quad (4)$$

Equation (4) applies only to same-class neighbor connections; all other entries are zero. Directed neighbor weights are first constructed and then symmetrized using $Z \leftarrow \max(Z, Z^T)$, with diagonal entries set to zero. Let D_g be the degree matrix, with $(D_g)_{ii} = \sum_j Z_{ij}$. The unnormalized graph Laplacian is $L_g = D_g - Z$. For the i -th row h_i of the hidden-layer output matrix, the graph Laplacian smoothness identity [13-15] gives:

$$\text{tr}(\beta^T H^T L_g H \beta) = \frac{1}{2} \sum_i \sum_j Z_{ij} \|h_i \beta - h_j \beta\|_2^2 \quad (5)$$

This term constrains output differences between same-class neighbors and preserves local consistency. The supervised graph uses only fitting-set labels; test labels are excluded from

graph construction and training. The constraint follows graph Laplacian smoothing [13-15], but no unlabeled samples are introduced in this study.

2.5. Joint Objective and Solution

To account for both class imbalance and local relationships within each class, the weighted classification loss and graph regularization term are combined in a single training objective. The joint solution fits the class labels while reducing prediction differences between neighboring samples of the same class.

Building on the weighted loss of WELM and graph Laplacian smoothing [12-14], the class-weighted error and graph smoothness term are combined to construct the LE-WELM objective:

$$\min_{\beta} \frac{1}{2} \|\beta\|_F^2 + \frac{C_1}{2} \|W^{\frac{1}{2}}(H\beta - T)\|_F^2 + \frac{C_1 C_2}{2} \text{tr}(\beta^T H^T L_g H \beta) \quad (6)$$

Here, $C_2 \geq 0$ is the graph regularization coefficient. Differentiating Eq. (6) with respect to β and setting the derivative to zero yields the linear system:

$$(H^T W H + \frac{I}{C_1} + C_2 H^T L_g H) \beta = H^T W T \quad (7)$$

Equation (7) follows directly from Eq. (6). Solving Eq. (7) gives β without explicitly computing a matrix inverse. When $C_2 = 0$, the model reduces to WELM; further setting $W = I$ gives the unweighted regularized ELM. The prediction rule is:

$$\hat{y} = \arg \max_{c \in \{0,1,2,3,4\}} [h(x)\beta]_c \quad (8)$$

The random hidden-layer mapping requires approximately $O(NdL)$ operations, the graph regularization matrix computations require approximately $O(NkL + NL^2)$, and the linear solution requires approximately $O(L^3)$, in addition to the cost of nearest-neighbor search. $L = 80$ is fixed in the controlled sampling experiments. Inference requires only the fixed feature mapping, hidden-layer mapping, and matrix multiplication, without reconstructing the supervised graph.

3. Experiments and Results

The experiments evaluate the proposed method's fault recognition performance under class imbalance. The data distribution, evaluation protocol, and sampling settings are introduced first. Classifier comparisons, ablation studies, and controlled sampling experiments are then used to analyze overall performance, the roles of individual components, and the method's limitations.

3.1. Experimental Setup

3.1.1. Dataset and Class Distribution

The number of samples in each class determines how much information the model can learn and helps explain the diagnostic challenges. The dataset and its class distribution are therefore introduced first to identify the fault classes with limited samples.

The dataset contains 6032 normalized vibration spectra from 11 ESPs: 4801 Normal, 576 Unbalance, 295 Faulty sensor, 70 Misalignment, and 290 Rubbing samples. The Normal class accounts for 79.59% of the data, whereas Misalignment accounts for only 1.16%, giving a sample-count ratio of approximately 68.59:1. Predicting every sample as Normal would yield an accuracy of approximately 79.59% over the entire dataset, illustrating the need for class-

balanced metrics. The sensor anomaly and mechanical fault labels together define a five-class classification task.

3.1.2. Evaluation Protocol and Parameter Settings

Model evaluation must consider both overall prediction accuracy and the recognition of faults with fewer samples. This section describes the data partitions, evaluation metrics, and parameter settings that provide a common basis for subsequent comparisons.

An 8×4 protocol with 32 folds and predefined training, validation, and test indices is used. This protocol does not imply complete separation of equipment identities. The downstream classifiers use five random seeds, 20260810 through 20260814, and results are reported as the mean and standard deviation over 160 runs. Because the repeated folds share data, these runs are not treated as 160 fully independent experiments.

For ELM and WELM, the search covers $L \in \{20, 40, 60, 80, 100\}$ and $C_1 \in \{0.01, 0.1, 1, 10, 100, 1000, 10000\}$. The graph constraint additionally considers $k \in \{3, 5, 7, 9, 11\}$, $t \in \{3, 5, 7\}$, and $C_2 \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Under the natural distribution, configurations are selected using validation results. The controlled sampling experiments use representative fixed parameters determined from the existing search results: $L = 80$, $C_1 = 100$, $C_2 = 0.01$, $k = 7$, and $t = 5$. Parameters are not retuned for each sampling setting.

The evaluation metrics are accuracy (Accuracy), macro-averaged F1 (Macro-F1), balanced accuracy (BAcc), and the geometric mean of class recalls (G-mean) [2,17-19]. Let the precision and recall of class c be P_c and R_c , respectively. The F1 score of that class is:

$$F1_c = \frac{2P_cR_c}{P_c + R_c} \quad (9)$$

Macro-F1 is the arithmetic mean of the F1 scores of the five classes [17]:

$$\text{Macro-F1} = \frac{1}{5} \sum_{c=0}^4 F1_c \quad (10)$$

BAcc is the arithmetic mean of the recalls of the five classes [18]:

$$\text{BAcc} = \frac{1}{5} \sum_{c=0}^4 R_c \quad (11)$$

G-mean is the fifth root of the product of the five class recalls. It is defined as zero if any class has zero recall [2,19].

$$\text{G-mean} = \left(\prod_{c=0}^4 R_c \right)^{1/5} \quad (12)$$

3.1.3. Controlled Sampling Settings

Experiments using only the original data distribution cannot readily distinguish the effects of class proportions from those of fault sample availability. The available numbers of Normal samples and selected fault samples are therefore adjusted separately, while the test set remains unchanged.

The first experiment retains all fault samples and downsamples only the Normal class without replacement. For a requested ratio $R \in \{10, 20, 50\}$, the target number of Normal samples in an individual training or validation subset is:

$$n_0^{\text{target}} = \min\{n_0, R \times \min(n_1, n_2, n_3, n_4)\} \quad (13)$$

Here, n_0 is the original number of Normal samples in that subset, and n_1, n_2, n_3 , and n_4 are the original numbers of Unbalance, Faulty sensor, Misalignment, and Rubbing samples, respectively. If there are insufficient Normal samples, all available Normal samples are retained. R denotes the requested ratio of Normal samples to the smallest fault class; the actual ratio of the largest to the smallest class across all classes need not equal R exactly.

The second controlled sampling experiment retains all Normal and Unbalance samples and applies a cap $K \in \{5, 10, 20, 40\}$ only to the Faulty sensor, Misalignment, and Rubbing classes. The training and validation subsets are sampled separately and then merged. The final number of fitting samples in class c is defined as:

$$n_c^{\text{fit}} = \min(n_{c,\text{train}}, K) + \min(n_{c,\text{val}}, K), c \in \{2,3,4\} \quad (14)$$

Here, $n_{c,\text{train}}$ and $n_{c,\text{val}}$ denote the numbers of class- c samples in the training and validation subsets before sampling, respectively. The final fitting set contains at most $2K$ samples for each capped class. When fewer samples are available, all existing samples are retained without duplication or synthesis. This setting is termed capped sampling of selected fault classes and differs from a standard K -shot setting in which each class contains exactly K samples in the final fitting set.

In both experiments, WELM and LE-WELM share the same sampling indices and seeds, and the test set remains unchanged. Class weights and the supervised graph are computed only from the final fitting set. Because sample compositions differ between the experimental groups, changes in scores cannot be attributed entirely to model structure.

3.2. Results and Analysis

Using the settings described above, the overall performance of different classifiers is first compared, followed by an analysis of individual model components. Changes in sample availability and recognition performance for specific fault classes are then examined to identify the improvements and remaining limitations.

3.2.1. Model Comparison under the Natural Distribution

Table 1. Performance of downstream classifiers under the natural distribution

Method	Accuracy	Macro-F1	BAcc	G-mean
ELM	0.8944±0.0194	0.6361±0.0411	0.6288±0.0351	0.5882±0.0673
WELM	0.9007±0.0174	0.7004±0.0429	0.7393±0.0344	0.7220±0.0700
Decision tree	0.8864±0.0147	0.6643±0.0316	0.6984±0.0358	0.6801±0.0433
MLP	0.9164±0.0134	0.7411±0.0387	0.7706±0.0344	0.7585±0.0417
SVM	0.9138±0.0130	0.7371±0.0355	0.7666±0.0341	0.7552±0.0400
LE-WELM	0.9180±0.0140	0.7500±0.0330	0.7810±0.0300	0.7750±0.0450

Note: Values are means ± standard deviations over 32 folds × 5 seeds. All methods use the same 1960-dimensional input features.

The first comparison evaluates the proposed method against other classifiers while retaining the original class distribution. All methods use the same input features and data partitions, and both overall accuracy and class-balanced metrics are considered.

Table 1 presents the results under the natural distribution. LE-WELM achieves an Accuracy of 0.9180, a Macro-F1 of 0.7500, a BAcc of 0.7810, and a G-mean of 0.7750. Compared with WELM, the latter three metrics increase by 0.0496, 0.0417, and 0.0530, respectively. Thus, adding local graph constraints within each class improves both overall accuracy and class-balanced performance.

ELM achieves an Accuracy of 0.8944 but a Macro-F1 of only 0.6361. Adding class weights increases Macro-F1 to 0.7004, indicating that frequency compensation improves class balance. LE-WELM exceeds MLP and SVM in Macro-F1 by 0.0089 and 0.0129, respectively.

3.2.2. Ablation of Class Weighting and Graph Constraints

The performance of the full model alone does not establish the usefulness of each component. The sources of performance changes are examined by removing or retaining individual components, including class weighting and graph regularization, and by assessing the effect of kernel group selection.

Table 2 presents the ablation results. Graph regularization alone yields a Macro-F1 of 0.7203, exceeding the 0.7004 obtained with class weighting alone, although the difference in BAcc is smaller. Combining both constraints gives a Macro-F1 of 0.7500 and a BAcc of 0.7810, supporting their complementary roles.

Table 2. Ablation of the classification layer and feature interface

Configuration	Macro-F1	BAcc	G-mean
No weighting or graph constraint	0.6361±0.0411	0.6288±0.0351	0.5882±0.0673
Class weighting only	0.7004±0.0429	0.7393±0.0344	0.7220±0.0700
Graph constraint only	0.7203±0.0330	0.7448±0.0342	0.7296±0.0430
Class weighting + graph constraint	0.7500±0.0330	0.7810±0.0300	0.7750±0.0450
Joint constraints without GS-WKS	0.7409±0.0347	0.7726±0.0348	0.7614±0.0425

Removing GS-WKS reduces Macro-F1 from 0.7500 to 0.7409, indicating a modest benefit from kernel group selection in these experiments. The feature dimension decreases from 2000 to 1960, a reduction of 2%. This stage therefore primarily serves feature selection rather than a substantial reduction in input dimensionality.

3.2.3. Capped Sampling of Selected Fault Classes

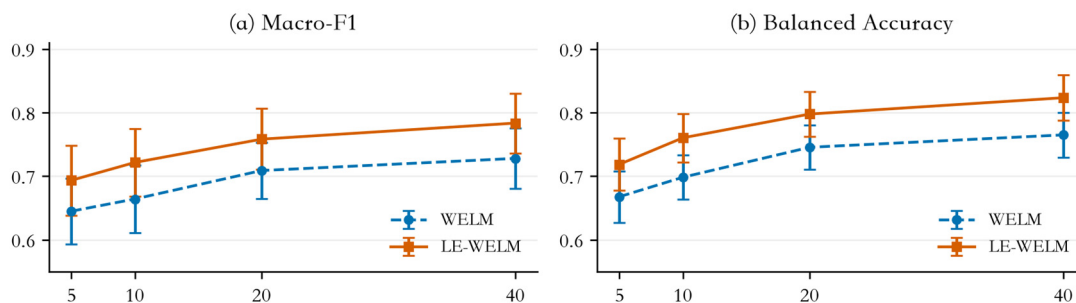


Fig 3. Classification performance under capped sampling of selected fault classes

When fault samples are scarce, the information available for learning is limited. Different sample caps are applied to selected fault classes to examine how recognition performance changes with sample availability and to compare the two methods under these conditions.

Figure 3 shows the capped-sampling results. Error bars denote the standard deviations across 32 folds and five random seeds. As K increases from 5 to 40, LE-WELM's Macro-F1 rises from 0.6936 to 0.7837, and its BAcc rises from 0.7188 to 0.8236. The corresponding metrics for WELM increase from 0.6450 and 0.6678 to 0.7281 and 0.7651, respectively. Both classifiers achieve better class-balanced performance as more samples become available in the capped fault classes.

The horizontal axis is the cap K on each selected fault class within an individual training or validation subset. At $K = 5, 10, 20$, and 40 , LE-WELM exceeds WELM in Macro-F1 by 0.0486, 0.0575, 0.0494, and 0.0556, respectively. Absolute performance remains relatively low at small K , showing that graph constraints cannot compensate fully for missing fault information.

3.2.4. Minority-class Recognition and Discussion

Aggregate metrics describe overall performance but do not directly show whether a specific fault is easier to recognize. Minority fault recall is therefore examined further. Statistical analysis and the experimental conditions are also considered when interpreting the results and the method's limitations.

Table 3 shows that Misalignment recall increases by 0.1138, 0.0702, and 0.1056 at $R = 20$, $K = 5$, and $K = 40$, respectively. Improvements in the aggregate metrics thus correspond to better recognition of some minority faults. However, recall remains below 0.5 at $K = 5$, and this class remains a diagnostic bottleneck.

Table 3. Misalignment recall under representative settings

Setting	WELM	LE-WELM	Mean difference
$R = 20$	0.6207 $\pm$ 0.1309	0.7345 $\pm$ 0.1144	0.1138
$K = 5$	0.4141 $\pm$ 0.1333	0.4843 $\pm$ 0.1325	0.0702
$K = 40$	0.5805 $\pm$ 0.1283	0.6861 $\pm$ 0.1345	0.1056

Note: K is the sample cap for each of the three selected fault classes in an individual training or validation subset. The final fitting set merges the two sampled subsets.

The paired mean differences in Macro-F1 are positive under all seven settings. The existing analysis used two-sided Wilcoxon signed-rank tests with Holm correction for 28 comparisons. Reference [20] discusses nonparametric tests and multiple comparisons, primarily for comparisons across datasets. Such procedures do not eliminate the dependence caused by shared data across repeated folds and multiple seeds in this study. The statistical results are therefore treated only as evidence of consistency under the current data partitions. The supervised graph depends on labels and local distances, and its neighborhoods may be insufficiently representative when very few samples are available. The controlled sampling experiments compare only WELM and LE-WELM and do not account for the labeling cost of the feature extraction stage. Future work should include additional baselines and experiments with limited labeling throughout the full pipeline.

4. Conclusion

This study develops LE-WELM to address class imbalance in ESP fault diagnosis using vibration spectra. The method operates on fixed features obtained through random wavelet kernels and kernel group selection. Inverse class-frequency weights balance the training contributions of different classes, while a same-class neighbor graph constrains the consistency of local sample outputs.

Under the compatible 32-fold evaluation protocol and the natural class distribution, LE-WELM achieves an accuracy of 91.80%, a Macro-F1 of 0.7500, and a balanced accuracy of 0.7810. The latter two metrics exceed WELM by 0.0496 and 0.0417, respectively. The ablation results indicate that class weighting and graph regularization are complementary.

In the controlled sampling experiments, LE-WELM improves Macro-F1 over WELM by 0.0486-0.0575. However, Misalignment recall remains low under small sample caps, indicating that graph constraints cannot replace missing fault information. Future work will adopt evaluation with strict separation of equipment identities and include additional comparison methods to further assess applicability across equipment and operating conditions.

References

- [1] Liu, R., Yang, B., Zio, E., et al. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mechanical Systems and Signal Processing*, 108, 33–47.
- [2] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- [3] Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232.
- [4] Guo, H., Li, Y., Shang, J., et al. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
- [5] Chawla, N. V., Bowyer, K. W., Hall, L. O., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- [6] Li, K., Li, S., Gao, X. Y., et al. (2025). Multivariate time-series tokenization Transformer and its application to fault diagnosis of electrical submersible pumps. *Control and Decision*, 40(4), 1145–1153. (in Chinese)
- [7] Mello, L. H. S., Oliveira-Santos, T., Varejão, F. M., et al. (2022). Ensemble of metric learners for improving electrical submersible pump fault diagnosis. *Journal of Petroleum Science and Engineering*, 218, 110875.
- [8] Varejão, F. M., Mello, L. H. S., Ribeiro, M. P., et al. (2024). An open source experimental framework and public dataset for vibration-based fault diagnosis of electrical submersible pumps used on offshore oil exploration. *Knowledge-Based Systems*, 288, 111452.
- [9] Feng, Y., Zheng, C., Chen, J., et al. (2025). Beyond deep features: Fast random wavelet kernel convolution for weak-fault feature extraction of rotating machinery. *Mechanical Systems and Signal Processing*, 224, 112057.
- [10] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks* (Vol. 2, pp. 985–990).
- [11] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3), 489–501.
- [12] Zong, W., Huang, G. B., & Chen, Y. (2013). Weighted extreme learning machine for imbalance learning. *Neurocomputing*, 101, 229–242.
- [13] Belkin, M., Niyogi, P., & Sindhvani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7, 2399–2434.
- [14] Huang, G., Song, S., Gupta, J. N. D., et al. (2014). Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44(12), 2405–2417.
- [15] von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 395–416.
- [16] Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems* 17 (pp. 1601–1608).

- [17] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- [18] Brodersen, K. H., Ong, C. S., Stephan, K. E., et al. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition* (pp. 3121–3124).
- [19] Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2–3), 195–215.
- [20] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.