International Journal of Science and Engineering Applications
Volume 14-Issue 11, 07 – 11, 2025, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1411.1003

Research on Interpretability of Bearing Fault Diagnosis Models for Transfer Learning

Qiuyue Wang
School of Electric Information
and Electrical Engineering
Yangtze University
Jingzhou, China

Abstract: With the widespread application of deep learning in fault diagnosis of mechanical equipment, its "black-box" nature has limited its deployment in safety-critical fields. Especially in transfer learning, where models need to adapt to different working conditions or equipment, the unobservability of their decision-making processes has exacerbated user distrust. Focusing on the interpretability of both the transfer process and the decision-making process, this paper proposes an interpretability analysis framework for deep learning-based bearing transfer fault diagnosis tasks, integrating the physical mechanisms of bearing faults. By analyzing the correlation between the transferable features of the model and the final diagnostic decisions, the internal logic of the transfer learning model in cross-condition diagnosis is revealed, thereby enhancing the transparency of the diagnostic process and the credibility of the results.

Keywords: Explainable Artificial Intelligence (XAI);Transfer Learning; Fault Diagnosis; Bearing; Deep Learning

1. Introduction

The health status of high-speed train bearings is crucial for ensuring driving safety. Data-driven intelligent diagnosis methods, especially deep learning, have achieved remarkable results in bearing fault diagnosis, but their decision-making process is usually like a "black box" and difficult to understand. When the diagnosis model needs to be transferred from the laboratory environment (source domain) to the actual train operating conditions (target domain), the model performance may be unstable due to differences in working conditions and loads. The unobservability of its transfer and decision-making processes further exacerbates the sense of distrust among engineers.[1]

Transfer learning is an effective means to address inter-domain distribution differences and achieve knowledge transfer. However, a transfer diagnosis model that only performs well but cannot be explained is difficult to be adopted in industrial sites that require high reliability. Therefore, the core of building a credible and transparent intelligent diagnosis system lies in breaking the "black box" of the model, that is, providing the model with full-process interpretability from feature construction and transfer process to final decision-making.

Focusing on the transfer fault diagnosis problem of high-speed train bearings, this paper aims to construct a system integrating multi-dimensional interpretability analysis. By combining physical priors, process quantification and post-hoc attribution, we not only pursue high diagnostic accuracy, but also strive to answer two key questions: "Why is the model credible?" and "What is the basis for diagnosis?" This provides a reliable path for the industrial application of intelligent diagnosis models.[2]

2. Construction of Multi-Dimensional Interpretable Transfer Diagnosis Model

The interpretable transfer diagnosis framework constructed in this study is shown in Figure 1, and its core consists of interpretable analysis from three dimensions: pre-event, in-process, and post-event.

In terms of pre-event interpretability, the research focuses on the physical meaning and traceability of the features themselves. Guided by the physical evolution mechanism of bearing faults, a feature system with clear physical connotations is built. To avoid the limitation that "black-box features" in traditional data-driven methods lack physical basis, the study accurately calculates four types of key theoretical characteristic frequencies based on the core geometric parameters of bearings and actual operating speeds using classical mechanical vibration theory formulas: Ball Pass Frequency Outer race (BPFO), Ball Pass Frequency Inner race (BPFI), Ball Spin Frequency (BSF), and Fundamental Train Frequency (FTF). This calculation process strictly follows the physical essence of bearing vibration—faulty components generate periodic impact or friction with other components during rotation, thereby exciting characteristic components of corresponding frequencies in the vibration signal—ensuring a one-to-one mapping between theoretical characteristic frequencies and actual fault modes.[3]

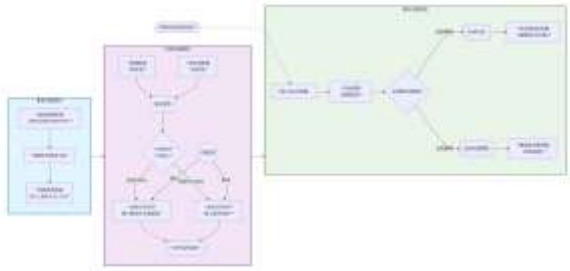

Figure 1: Multi-Dimensional Interpretable Transfer Diagnosis Framework

After obtaining the key theoretical characteristic frequencies, to further quantify the representation intensity of fault information in the vibration signal, the study uses envelope analysis as technical support to construct energy quantification indicators based on the harmonic frequency bands of theoretical characteristic frequencies. Envelope demodulation is performed on the bearing vibration signal, and specific frequency bands are constructed around BPFO, BPFI, BSF, FTF, and their respective harmonics. The vibration energy values within each frequency band are obtained and defined as harmonic frequency band energy ratio indicators. Meanwhile, energy normalization is introduced to calculate derived indicators such as the ratio of energy in each fault frequency band to the total signal energy and the energy ratio between different fault frequency bands, realizing the robust representation of fault features under complex operating conditions. This physically mechanism-based feature system directly maps the physical evidence of specific fault modes in the vibration signal, providing a solid and traceable prior knowledge anchor for the model’s subsequent learning and decision reasoning, and ensuring the consistency between the model’s decision-making process and the physical evolution law of bearing faults from the source.

For in-process interpretability, the research aims to solve the "knowledge transfer black-box" problem by constructing an interpretability framework from two aspects: the transparent design of distribution alignment methods and multi-dimensional quantitative analysis of alignment effects. A strategy combining Correlation Alignment (CORAL) and Class-Conditional Correlation Alignment (C-CORAL) is adopted. By matching the second-order covariance statistics of features from the source domain and target domain, accurate alignment of inter-domain feature distributions is achieved.[4]The interpretability of this strategy lies in its explicit linear transformation process: through the construction of an analyzable linear transformation matrix, the source domain feature space is mapped to a new space highly matching the second-order statistical characteristics of the target domain. The entire mapping process completes the matching of covariance statistics through clear matrix operations, ensuring process transparency. To accurately evaluate the alignment effect, the study introduces a covariance gap decomposition mechanism. By calculating the Frobenius norm of the feature covariance matrices of the source and target domains before and after alignment, the inter-domain covariance difference is converted into a quantifiable numerical indicator. Matrix decomposition technology is used to decompose this overall covariance gap into the variance gap of each individual feature and the covariance gap between different features, thereby quantitatively verifying the effectiveness of the alignment strategy from a statistical perspective.

Considering the potential class-level differences in global distribution alignment, the study further introduces class-conditional Maximum Mean Discrepancy (MMD) as a supplementary metric for class-level alignment effects. This metric calculates the distance between the feature distributions of the same fault class samples in the source and target domains in the Reproducing Kernel Hilbert Space (RKHS), enabling accurate quantification of class-level inter-domain differences.[5]This fine-grained class-based difference measurement not only intuitively reveals the class-level alignment capability of the C-CORAL method but also provides a clear direction for the targeted improvement of subsequent alignment strategies, improving the evaluation system for the interpretability of the transfer process.

In terms of post-event interpretability, the research delves into the model’s decision results to reveal the underlying logic of the model’s specific diagnostic decisions from both micro and macro dimensions. For the micro-level analysis of single-sample decision logic, the SHAP (SHapley Additive exPlanations) attribution method based on game theory is introduced. By quantifying the marginal contribution of each feature in all possible feature subset combinations, a fair and unique SHAP value is assigned to each feature.[6]This value can accurately characterize the contribution degree and direction of the feature to the prediction result. Through the visualization of SHAP values, the key driving features and inhibitory factors for the diagnostic result of each sample can be intuitively traced, realizing "transparent traceability" of single-sample decision logic.

To further grasp the overall decision rules of the model in the target domain from a macro perspective, based on the model’s prediction results for all samples in the target domain, the study constructs a sparse logistic regression model with L1 regularization as a global surrogate model. This surrogate model takes the input features of the original model as independent variables and the diagnostic results of the original model as dependent variables. Through the sparsity constraint of L1 regularization, it automatically selects the features most critical to the model’s global decision-making. The regression coefficients of the surrogate model have clear physical meanings: the absolute value of the coefficient reflects the dominant role intensity of the feature in the global diagnostic decision, while the sign of the coefficient reveals the correlation direction between the feature and a specific fault class. The construction of the global surrogate model abstracts the complex decision-making process of the original model into an interpretable linear relationship, clearly summarizing the model’s "global decision logic" and providing macro-level support for the reliability evaluation and generalization ability analysis of the model in cross-domain scenarios.

3. Experimental Verification and Interpretability Analysis

3.1 Inter-domain Distribution Alignment: Interpretability Verification

To systematically verify the inter-domain feature alignment effectiveness of the Class-Conditional Correlation Alignment (C-CORAL) method, this study establishes a dual verification system that combines visualization and quantitative analysis. Through t-SNE feature visualization and covariance discrepancy decomposition, the optimization effect of C-CORAL on the feature distributions of the source domain and target domain is fully revealed.

In terms of visualization analysis, the t-SNE algorithm is used to map high-dimensional features to a two-dimensional space. As shown in Figure 2-a, before alignment, the samples of the source domain and target domain exhibit an obvious "cluster separation" phenomenon with clear inter-domain boundaries, indicating significant distribution differences in the original feature space. After C-CORAL alignment (Figure 2-b), the samples of the two domains form an interleaved distribution pattern, and the inter-domain boundaries basically disappear. This intuitively proves that the method can effectively reduce the inter-domain distribution discrepancy.

At the quantitative analysis level, the evaluation based on covariance discrepancy decomposition shows that the total covariance discrepancy between the source domain and the target domain decreases sharply from 1.895×10^{6} before alignment to 2.183×10^{5}, with a reduction rate of over 88%. This confirms the effectiveness of C-CORAL from a global statistical perspective. Further feature-level decomposition results (Figure 3) indicate that the corresponding covariance discrepancies of both time-domain statistical features and physical mechanism features are significantly reduced. Notably, the covariance discrepancy of mechanism-based features decreases particularly obviously, which proves that while unifying the statistical distribution, C-CORAL effectively preserves and enhances the physical consistency of key fault features.

This dual verification not only confirms the effectiveness of the C-CORAL method but also reveals its unique advantages: in the process of unifying the inter-domain statistical distribution, it not only does not destroy the physical information of fault features but also enables the physical representations of the same fault mode to show higher consistency across different domains through feature alignment. This provides an important guarantee for subsequent diagnostic models to accurately capture the essence of faults.

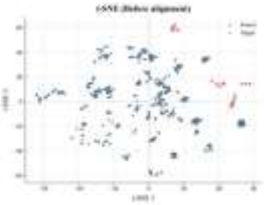

Figure 2-a: t-SNE Distribution Representation of Samples Before Alignment

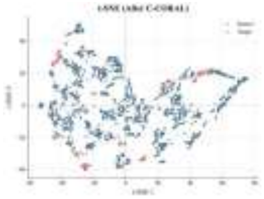

Figure 2-b: t-SNE Distribution Representation of Samples After Alignment

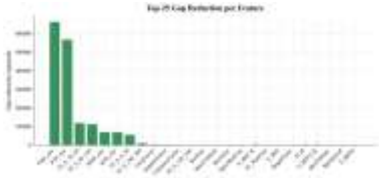

Figure 3: Visualization of Top 25 Gap Reductions

3.2 Analysis of Changes in Intra-Class Distribution Discrepancies

In cross-domain fault diagnosis, the effectiveness of global distribution alignment is not entirely equivalent to the quality of class-level alignment. If only the overall inter-domain feature distribution is matched, but significant discrepancies still exist between samples of the same fault class in the source domain and target domain, the model's accurate identification of fault classes in the target domain will be directly affected. To address this, this study quantitatively analyzes the alignment capability of the C-CORAL method from a fine-grained class perspective by calculating the Class-intra Squared Maximum Mean Discrepancy (Class-intra MMD²). The results, as shown in Figure 4, clearly demonstrate the method's significant effectiveness in narrowing the inter-domain distribution gap of samples with the same fault class.

The core value of the Class-intra MMD² metric lies in its ability to accurately measure the distribution distance of samples belonging to the same fault class between the source domain and target domain in the feature space. A larger distance value indicates more significant feature differences of the same fault class between the two domains, making it harder for the model to achieve cross-domain generalization; conversely, a smaller value means higher feature consistency of the same fault class, which provides favorable conditions for accurate classification. From the comparison results in Figure 4, before alignment using C-CORAL, the Class-intra MMD² value for ball bearing faults (denoted as Class B) is as high as 1.44, and that for outer race faults (denoted as Class OR) is 0.93. This data indicates that even for the same fault mode, there is still an obvious shift in its feature expression between the source domain and target domain. For example, the vibration energy distribution characteristics and fault frequency harmonic structure exhibited by ball bearing fault samples in the source domain are significantly different from the corresponding features of samples with the same fault in the target domain. Such differences can cause the model to misclassify ball bearing faults in the target domain as other classes.

After alignment using the C-CORAL method, the intra-class inter-domain discrepancies of both fault classes are significantly reduced: the Class-intra MMD² value for ball bearing faults (Class B) decreases from 1.44 to 0.63, a reduction of over 56%; the Class-intra MMD² value for outer race faults (Class OR) drops from 0.93 to 0.17, a substantial reduction of 82%. [7]This change is not only reflected in a significant numerical decrease but also reflects the C-CORAL method's precise optimization of class features. Through covariance alignment under class-conditional constraints, the method not only matches the global statistical properties between domains but also focuses more on optimizing the feature consistency of samples with the same fault class. This enables highly consistent matching of key physical features of ball bearing faults and outer race faults in the source domain and target domain—such as the energy proportion in the BSF (Ball Spin Frequency) band corresponding to ball bearing faults and the BPFO (Ball Pass Frequency Outer race) harmonic distribution corresponding to outer race faults.

The above comparison results of Class-intra MMD² strongly confirm the alignment effectiveness of the C-CORAL method from a fine-grained perspective: it does not simply narrow the overall inter-domain discrepancy but can achieve "precision alignment" for different fault classes, effectively eliminating the feature shift problem of the same fault class in cross-domain scenarios. This high-quality class-level alignment lays a solid foundation for the subsequent fault diagnosis model's classification tasks in the target domain. When the model processes fault samples in the target domain, it can make decisions based on feature expressions that are highly consistent with those of the same fault class in the source domain, greatly reducing the risk of misjudgment caused by inter-domain discrepancies and ultimately improving the accuracy of cross-domain fault diagnosis.

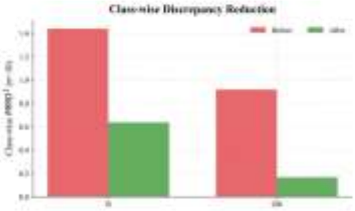

Figure 4: Comparison of Class-intra MMD²

3.3 Decision Logic Interpretation Based on SHAP

To address the "decision black box" issue of the RBF-SVM model, this study employs the SHAP method for post-hoc interpretability analysis. Through two dimensions—class-level feature importance and fine-grained single-feature attribution—it systematically reveals the intrinsic logic behind the model's decisions.

In the class-level feature importance analysis, the study finds that there are significant differences in the core discriminative features across different fault classes. [8]As shown in Figure 5, the identification of ball bearing faults (Class B) mainly relies on the time-frequency energy feature in the low-frequency band (TF_E_0_30), which is consistent with the physical characteristic of ball bearing defects inducing low-frequency impacts. For the diagnosis of inner race faults (Class IR) (Figure 6), the peak-to-peak value of the original vibration signal (PTP_raw) emerges as the most discriminative feature, reflecting the severe amplitude fluctuations caused by inner race faults rotating with the shaft. This targeted matching relationship between features and fault modes confirms that the model can adaptively select discriminative bases according to the dynamic characteristics of faults.

At the fine-grained analysis level, the SHAP beeswarm plot (Figure 8) reveals the micro-level causal relationship between feature values and prediction results. Taking PTP_raw as an example, samples with high feature values (in blue) mostly correspond to positive SHAP values, driving the model to make a fault judgment; in contrast, samples with low feature values (in red) correspond to negative SHAP values, suppressing fault prediction. This pattern is fully consistent with the physical mechanism of faults: inner race damage leads to an increase in vibration amplitude, while normal operation maintains a low amplitude level. Similarly, samples with high values of features such as TF_E_0_30 also exhibit positive contributions, further verifying the indicative role of low-frequency energy in ball bearing fault diagnosis.

This multi-level interpretability analysis transforms the abstract model decisions into a traceable causal chain. It not only verifies the physical rationality of the model's decisions but also provides a direct basis for evaluating the reliability of diagnostic results.

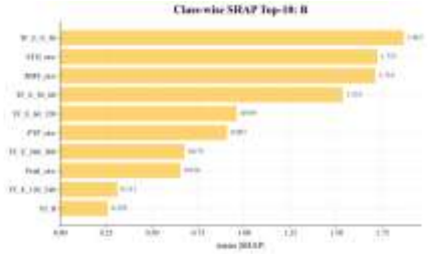

Figure 5: SHAP Value Features for Ball Bearing Faults

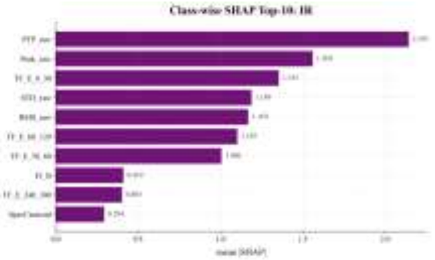

Figure 6: SHAP Value Features for Inner Race Fault Class

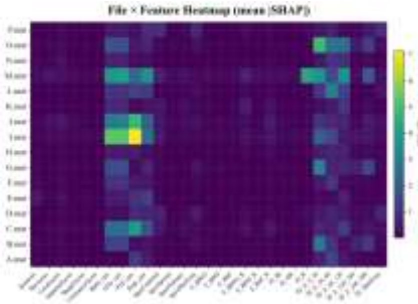

Figure 7: Average SHAP Values of Files and Feature Dimensions

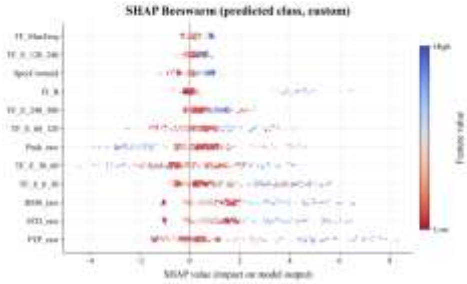

Figure 8: SHAP Beeswarm Plot for Custom Prediction Classes

3.4 Interpretability Verification of Global Decision Logic

To extract the decision logic of the RBF-SVM model from a global perspective and avoid the fragmented limitations of local interpretations, this study constructs a global sparse logistic regression surrogate model (results shown in Figure 9). Through the significance analysis of the model's regression coefficients, it systematically summarizes the core dependent features and decision rules in the diagnosis of different fault classes, further verifying the consistency between the model's decisions and the physical mechanisms of faults.

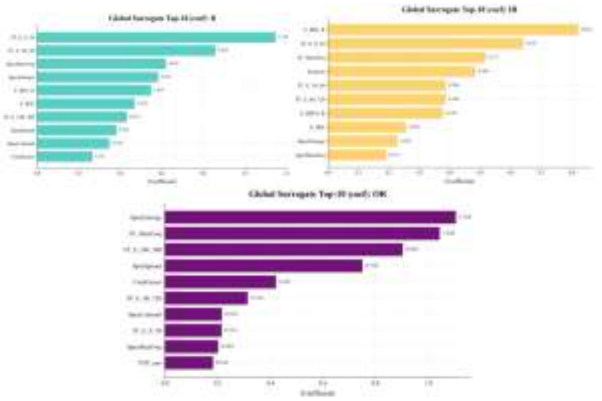

Figure 9: Visualization of Global Surrogate Features

From the results of the surrogate model, it can be observed that the model exhibits distinct feature dependency differences in distinguishing between different fault classes:

For ball bearing faults (Class B), the diagnostic decision mainly relies on time-frequency energy features in the low-frequency band. [9]Meanwhile, mechanism-based features related to ball bearing faults (fault index, frequency band energy ratio) provide a stable positive contribution to the decision, which is consistent with the dynamic characteristics of low-frequency impacts caused by ball bearing faults.

For inner race faults (Class IR), the diagnostic process is dominated by mechanism-based features and , supplemented by energy features in the high-frequency band to form a multi-dimensional criterion. This aligns with the physical nature of inner race faults—rotating at high speed with the main shaft, they exhibit both mechanism-specific characteristics and high-frequency vibrations.

In the discrimination of outer race faults (Class OR), the signs of the regression coefficients for core mechanism-based features and fully match the feature evolution law of outer race faults. Additionally, the reduction in high-frequency components is identified by the model as a key discriminative clue, which is consistent with the characteristic of outer race faults where vibration energy is concentrated at specific mechanism-related frequencies.

This global decision pattern is mutually corroborated by the local interpretations from SHAP. Together, they demonstrate that the model's decision logic is consistently tied to the physical mechanisms of bearing faults, forming a unified explanatory loop from the macro to the micro level. This fundamentally verifies the credibility of the model.

4. Conclusion

To address the issue of insufficient credibility caused by the "black-box" model decision-making in cross-domain transfer diagnosis of high-speed train bearings, this study constructs a three-dimensional interpretability analysis framework covering pre-hoc feature design, transfer process optimization, and post-hoc decision interpretation. Through systematic experimental verification and quantitative analysis, the following core conclusions are drawn:

At the pre-hoc interpretability level, the feature system constructed based on the physical evolution mechanism of bearing faults (such as theoretical frequencies like BPFO and BPFI, and the corresponding harmonic frequency band energy ratio provides the transfer diagnosis model with prior knowledge anchors with clear physical meanings. These features directly map the causal relationship between fault modes and vibration signals, avoiding the limitation of statistical correlation of purely data-driven features from the source, and laying a solid physical foundation for credible diagnosis.

At the transfer process interpretability level, inter-domain feature distribution alignment is achieved through global CORAL and class-conditional CORAL (C-CORAL). Combined with covariance discrepancy decomposition and intra-class MMD² quantitative analysis, a dual-guarantee mechanism of "transparent alignment process + quantitative effect evaluation" is formed. The reduction of total covariance discrepancy by more than 88% and the significant decrease in intra-class MMD² not only intuitively demonstrate the effective reduction of inter-domain distribution differences but also clearly identify the alignment effect of key fault features, ensuring the transfer process is traceable and verifiable.

At the post-hoc interpretability level, the combination of SHAP attribution analysis and the global sparse logistic regression surrogate model enables a comprehensive interpretation of the model's decision logic from local to global perspectives: SHAP values reveal the positive and negative contribution directions of features in single-sample diagnosis, while the surrogate model extracts the core discriminative rules for different fault classes (e.g., ball bearing faults rely on low-frequency energy, outer race faults focus on the feature). These two aspects mutually corroborate, fully confirming the high consistency between the model's decisions and the dynamic characteristics of bearing faults.

In summary, the three-dimensional interpretability framework proposed in this study significantly improves the transparency and credibility of the transfer diagnosis model for high-speed train bearings, effectively solving the problems of "knowledge transfer black box" and "ambiguous decision logic" in cross-domain diagnosis. It not only provides technical references for interpretability research in the field of bearing fault diagnosis but also offers important theoretical support and practical examples for the reliable implementation and engineering application of intelligent diagnosis systems in complex industrial scenarios.

5. References

[1] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

[2] Wen, L., Gao, L., & Li, X. (2017). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(1), 136-144.

[3] Sun, B., Saenko, K. (2016). Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In European Conference on Computer Vision (ECCV).

[4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).

[5] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub.

[6] Zhang, W., Li, X., Jia, X. D., Ma, H., Luo, Z., & Li, X. (2020). Machinery fault diagnosis with imbalanced data using deep transfer learning. Measurement, 151, 107219.

[7] Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587.

[8] Wang, J., Li, S., An, Z., Jiang, X., & Yang, B. (2021). A physical model-informed transfer learning framework for bearing fault diagnosis with small data. Mechanical Systems and Signal Processing, 155, 107634.

[9] Grezmak, J., Zhang, J., Wang, P., Loparo, K. A., & Gao, R. X. (2019). Interpretable convolutional neural network for bearing fault diagnosis. IEEE Transactions on Instrumentation and Measurement, 69(5), 2180-2190