IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 April 2024, accepted 28 June 2024, date of publication 17 July 2024, date of current version 5 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430010

SURVEY

Enhancing Reliability Through Interpretability: A Comprehensive Survey of Interpretable Intelligent Fault Diagnosis in Rotating Machinery

GANG CHEN[1,2], **(Member, IEEE)**, **JUNLIN YUAN**[1], **YIYUE ZHANG**[1], **HANYUE ZHU**[1], **RUYI HUANG**[1], **(Member, IEEE)**, **FENGTAO WANG**[3], **AND WEIHUA LI**[4], **(Senior Member, IEEE)**
[1]Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 511442, China
[2]The State Key Laboratory of Mechanical Transmissions, Chongqing University, Chongqing 400044, China
[3]Department of Mechanical Engineering, College of Engineering, Shantou University, Shantou 515063, China
[4]School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, China

Corresponding author: Gang Chen (gangchen@scut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 52305105, in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2022A1515240027 and Grant 2023A1515010812, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1582, and in part by the Open Fund of State Key Laboratory of Mechanical Transmissions under Grant SKLMT-MSKFKT-202321.

ABSTRACT This paper presents a comprehensive survey on interpretable intelligent fault diagnosis for rotating machinery, addressing the challenge of the ‘‘black box’’ nature of machine learning techniques that hampers reliability in automated diagnostic processes. It underscores the growing importance of interpretability in intelligent fault diagnosis (IFD), marking a shift from traditional signal processing methods to machine learning-based approaches that necessitate transparency for trustworthiness. Our review systematically collates and examines the spectrum of interpretability in IFD, distinguishing between post-hoc and ante-hoc strategies. We detail mainstream post-hoc methods, their applications, and critique their limitations, particularly the absence of physical significance. The survey then explores ante-hoc methods that incorporate physical knowledge upfront, enhancing interpretability. By categorizing and evaluating three distinct knowledge embedding approaches, we shed light on their unique applications. Conclusively, we highlight emerging research directions and challenges in the field, aiming to equip readers with a nuanced understanding of current methodologies and inspire future studies in making IFD more reliable and interpretable.

INDEX TERMS Intelligent fault diagnosis, post hoc interpretation, ante hoc interpretation, explainable artificial intelligence, deep learning, rotating machine.

I. INTRODUCTION

Rotating machinery, critical to industries such as wind energy, aerospace, and maritime, plays a foundational role in modern manufacturing [1]. These machines, including wind turbine transmission systems, helicopter gearboxes, and marine propulsions, operate in challenging environments, making failures inevitable and posing risks of accidents and economic losses [2]. The importance of accurate fault diagnosis (FD) for these machines cannot be overstated, as it enables timely maintenance, extends service life, and ensures operational safety [3], [4], [5].

Traditionally, FD in rotating machinery has relied on the extensive experience and expert knowledge of engineers skilled in data signal processing. Engineers could identify malfunctions through auditory cues or by analyzing vibration

The associate editor coordinating the review of this manuscript and approving it for publication was Mehrdad Saif.

signals across the time, frequency, and time-frequency domains. Techniques such as spectral kurtosis [6], resonance demodulation, [7] empirical mode decomposition [8], variational mode decomposition [9], and wavelet transform [10] have been effectively employed for this purpose. Each method has demonstrated significant success in diagnosing faults. However, these traditional methods suffer from a crucial limitation: the need for specific feature extraction tailored to each unique fault type [11], [12]. This specificity makes the process cumbersome and less applicable in a general context, highlighting a gap for more universally applicable diagnostic approaches [7].

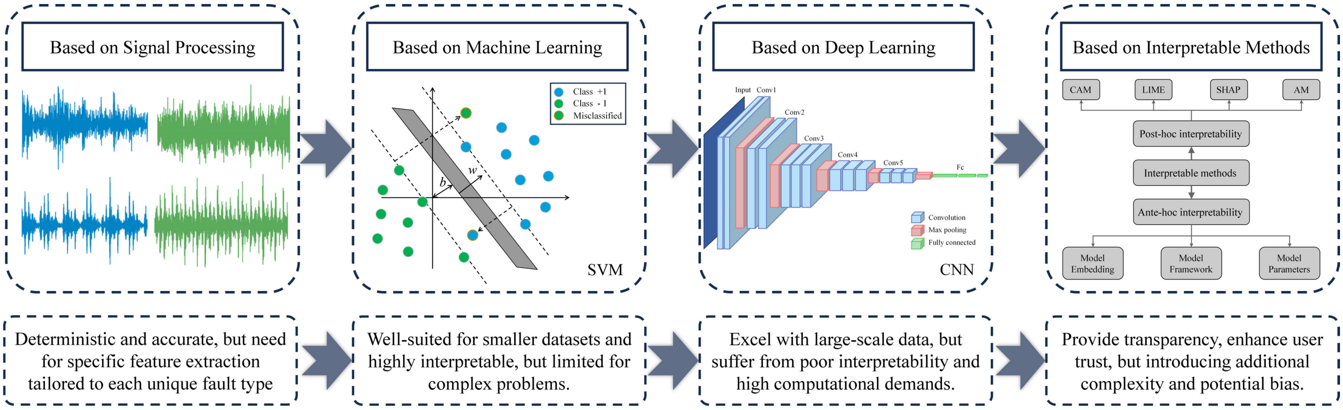
The evolution of sensor technology and the exponential growth of monitoring data have catalyzed the development of intelligent fault diagnostic (IFD) methodologies [8]. These methodologies leverage machine learning (ML) models to interpret various monitoring signals and ascertain the health status of machinery, showcasing a shift towards automation in fault diagnosis [9]. For instance, Hidden Markov Models have been utilized for their strength in modeling dynamic time series and classification capabilities in rotating machinery's acceleration and deceleration processes [10]. Similarly, the integration of wavelet packet decomposition with Principal Component Analysis and support vector machines (SVM) for diagnosing faults in bearings represents the broader application of ML in this domain [13]. Additionally, other ML models, such as k-nearest neighbour [14], and artificial neural networks [15], have been widely applied in the realm of IFD, further illustrating the diverse and expanding applications of ML in fault diagnosis.

Abbr.	Meaning.
IFD	Intelligent Fault Diagnosis.
IIFD	Interpretable Intelligent Fault Diagnosis.
FD	Fault Diagnosis.
SVM	Support Vector Machines.
ML	Machine Learning.
DL	Deep Learning.
NLP	Natural Language Processing.
CNNs	Convolutional Neural Networks.
CAM	Class Activation Mapping.
LIME	Local Interpretable Model-agnostic Explanations.
SHAP	SHapley Additive exPlanations.
AM	Attention Mechanism.
CV	Computer Vision.
PHM	Prognostics and Health Management.
CWRU	Case Western Reserve University.
EDM	Electrical Discharge Machining.
PU	Paderborn University.
FCF	Fault Characteristic Frequency.
XAI	Explainable Artificial Intelligence).
RF	Random Forest.
LRP	Layer-Wise Relevance Propagation.
GA	Gradient Ascent.
CIU	Contextual Importance and Utility.
LGSC	Layered General Sparse Coding.
NISTA	Iterative Soft Thresholding Algorithm.
MCAN	Multi-Scale Component Analysis Network.
MCA	Morphological Component Analysis.
SPINN	Signal Processing Informed Neural Network.
TLNN	Temporal Logic Neural Network.
STL	Signal Temporal Logic.
RSFDS	Restricted Sparse Frequency-domain Space.
RL	Reinforcement Learning.

In recent years, deep learning (DL) techniques have emerged as a frontier for automating and enhancing the precision of data analysis, gaining acclaim in fields such as natural language processing (NLP) and image classification [16]. DL's ability to autonomously extract features and facilitate end-to-end fault diagnosis represents a significant leap over traditional ML-based IFD methods, which require extensive manual effort and expertise for feature extraction. Among various DL models, convolutional neural networks (CNNs) and attention mechanisms have been particularly notable for their achievements in computer vision and NLP. Their application in fault diagnosis has seen an upsurge in comprehensive research, offering improved reliability and generalization capabilities [17], [18], [19], [20].

Despite their advanced diagnostic capabilities, the increasing reliance on DL models has heightened the need for interpretability. The "black box" nature of these models often results in a lack of transparency regarding their decision making processes, undermining user trust. To enhance the transparency of Deep Learning (DL) models, Explainable Artificial Intelligence (XAI) techniques have been extensively studied and can broadly be divided into post-hoc and ante-hoc interpretable methods. Compared to the aforementioned methods, XAI maintains the transparency of traditional signal processing and machine learning methods while inheriting the capability of DL to handle large datasets. The development and differences of fault diagnosis methods are illustrated in Figure 1.

Existing reviews on IFD, such as those by Liu et al. [21], Lei et al. [7], Zhang et al. [22], Lv et al. [23], and Zhu et al. [24], have offered comprehensive insights into the applications of traditional ML and the evolution of IFD methodologies, including the advent of DL. However, these reviews do not adequately address the critical aspect of interpretability within interpretable intelligent fault diagnosis (IIFD). Interpretability, as highlighted by Lei et al. [7] and Zhu et al. [24], is essential for the industrial application of IFD models, enabling users to understand and trust the model's predictions. Despite its importance, there is a notable gap in the literature regarding comprehensive reviews focused specifically on interpretability within the IIFD domain. This gap is significant, as interpretability not only enhances trust by clarifying the decision-making process but also facilitates the identification of relevant parameters used for classification, improving model evaluation [25]. Without a clear basis for the model decision, users may



Deterministic and accurate, but need for specific feature extraction tailored to each unique fault type

Well-suited for smaller datasets and highly interpretable, but limited for complex problems.

Excel with large-scale data, but suffer from poor interpretability and high computational demands.

Provide transparency, enhance user trust, but introducing additional complexity and potential bias.

FIGURE 1. The development and differences of IFD.

find the outcomes less credible. However, there is neither a clear mathematical definition of interpretability nor a complete theory and method system of pure mathematical analysis [26]. Hence, we believe it is necessary to review and summarize the current interpretability methods and their applications in IFD, and to highlight the challenges they face.

Our survey seeks to address this gap by providing a detailed review of current interpretability methods in IFD, exploring their applications, and discussing the challenges they face. By highlighting the importance of developing interpretable models, this review aims to bridge the existing gap between sophisticated diagnostic capabilities and their practical industrial application, building upon the foundation laid by previous literature and guiding future research toward more transparent and trustworthy IFD models. In this paper, we summarize the current research on IIFD, offering a reference for future studies in this direction. The contributions of this review are articulated as follows:

1) This review systematically categorizes IIFD approaches into two primary categories: ante-hoc and post-hoc interpretations. Ante-hoc interpretation involves proactive modifications to the network architecture or training process to enhance interpretability and transparency. In contrast, post-hoc interpretation focuses on elucidating the decision-making processes of already trained neural networks. We provide comprehensive definitions, detailed explorations of various interpretable methods, and summarize their applications and inherent limitations.

2) While post-hoc interpretation has received considerable attention and yielded numerous advancements in the last two years, research on ante-hoc interpretation is relatively nascent. Our in-depth examination of the current challenges in ante-hoc IIFD uncovers promising research avenues that warrant further exploration.

3) Our survey is at the forefront of collating approaches to ante-hoc IIFD, organizing them into three novel categories: interpretability of model embedding, model framework, and model parameters.

This classification aims to address the existing research void in ante-hoc interpretation and spark innovation in developing fault diagnosis methodologies that are both effective and transparent. We intend to guide researchers in this domain toward uncovering new opportunities and contributing to the ongoing progress of IFD in the industrial realm.

The rest of this review is organized as follows. In section II, we summarize the research methodology, dataset, and the initial analysis of collected papers. Section III reviews the applications of post-hoc interpretation, which is considered a method that allows users to know better how algorithms make decisions. Section IV argues applications of ante-hoc interpretation to IFD including the motivation, the definitions, and some exploratory works. In Section V we further display a prospect when combined with the challenges of IIFD. Conclusions are enclosed in Section VI.

II. RESEARCH METHODOLOGY AND INITIAL ANALYSIS

A. RESEARCH METHODOLOGY

This review meticulously evaluates the evolving field of IIFD, shedding light on significant contributions that enrich both academic research and industrial practice while pinpointing prevailing research gaps. To ensure an exhaustive overview, our literature survey spanned publications from 2017 to March 2024, drawing from an array of esteemed scientific databases including Science Direct, IEEE Xplore, Springer, Scopus, and Web of Science. To incorporate the cutting-edge advancements in the field, we also reviewed preprints from arXiv. Employing a strategic selection of keywords and screening criteria led to the identification of 205 papers that are central to the theme of IIFD. This curated collection serves as a foundational resource for those delving into the intricacies of current trends, methodologies, and challenges in IIFD research.

To navigate the extensive realm of interpretable deep learning within fault diagnosis, we crafted a comprehensive three-level keyword tree, illustrated in Figure 2. This structured methodology facilitated a deep dive into the world of interpretable deep learning, categorizing it into post-hoc and ante-hoc interpretability. A focused search that paired ‘‘fault diagnosis’’ with terminology related to post-hoc interpretability methods, which involves techniques applied after model training to clarify how decisions are made, including Class Activation Mapping (CAM), Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Attention Mechanism (AM), revealed an impressive corpus of 175 publications. This exploration not only highlighted the diversity of post-hoc explanation techniques within the IIFD domain but also underscored the prolific research activity in this area.

Recognizing the nascent stage of ante-hoc interpretability in IFD — characterized by a lack of consensus on methods — we employed a targeted keyword strategy encompassing ''ante-hoc/intrinsic + explainable/interpretable + fault diagnosis.'' This led to a novel classification of ante-hoc explanation methods into three innovative categories:

1) Model Embedding Interpretability: refers to the integration of interpretable components or structures within a model to enhance its transparency and the ability to provide clear explanations for its predictions.
2) Model Framework Interpretability: refers the capacity of a machine learning framework to clearly explain its underlying logic and decision-making pathways.
3) Model Parameters Interpretability: refers to the ability to understand and explain the relationship between the parameters of a predictive model and the phenomena or outcomes it is designed to diagnose.

An advanced keyword search integrating ''fault diagnosis'' with terms such as ''physically informed,'' ''algorithms unrolling,'' ''interpretable kernel,'' ''sparse,'' ''formal language,'' and ''framework'' unearthed over 30 pertinent publications. This effort not only enhanced our understanding of ante-hoc interpretability's application in IIFD but also contributed to filling the knowledge gap in this emerging area.

However, the process of conducting a literature review in a rapidly advancing field like IIFD, especially one involving explainable deep learning technologies, is inherently fraught with challenges. One notable issue was the potential overlook of relevant studies, possibly due to our dependence on a specific set of keywords. This approach might have missed capturing the entire spectrum of ongoing research. For example, several studies employ techniques like CAM and AM with a primary focus on assessing model performance rather than interpretability per se. Moreover, the aspect of ''uncertainty'' as an integral component of model interpretability — although acknowledged by researchers [1], [27] — was not incorporated as a keyword in our search strategy. This oversight might have resulted in excluding studies where ''uncertainty'' plays a crucial role in interpretability or where interpretability is discussed in an implicit context. This experience underscores the critical importance of adopting a flexible and comprehensive literature review strategy, capable of adapting to the nuances and rapid developments within such technologically dynamic domains.

B. INITIAL ANALYSIS

We conducted a preliminary analysis of the collected literature. Firstly, we analyzed the literature based on publication year and keywords to summarize its development trends and hotspots. Then, by examining the research subjects and the main problems addressed, we summarized the applications of the methods discussed in the literature. Finally, we classified the explainable technologies in rotating machinery IFD by drawing analogies to the traditional fault diagnosis process.

1) TRENDS AND HOTSPOTS

In the early stages of ML, models like expert systems and decision trees were inherently interpretable, enabling users to easily understand their decision-making processes. However, the advent of DL marked a paradigm shift towards models with complex, ''black box'' architectures, making their decision mechanisms challenging to decipher. This transition has underscored an urgent need for interpretability and transparency in DL, a demand that has grown increasingly vital across various domains over the last decade. By 2023, research into DL interpretability had made significant strides, particularly within Computer Vision (CV) and NLP. These fields have seen the development of several general post-hoc explanation tools, such as CAM, LIME, and SHAP, which have greatly enhanced our understanding of DL models. Research on IIFD began to gain momentum around 2017 and has seen rapid development since. Mechanical fault diagnosis, compared to CV and NLP, faces a more urgent need for deep learning interpretability due to the stringent reliability and stability requirements in the industrial sector. The ''black box'' nature of DL models poses a significant challenge for their application in fault diagnosis, underscoring the vital importance of IIFD development for their effective and trustworthy implementation in critical industrial operations.

Drawing from the comprehensive review by Lei et al. [7], which summarized IFD research up to 2019, we have further analyzed recent trends in IIFD literature, as illustrated in Figure 3. This analysis highlights a marked increase in interest in IIFD from 2019 onward, coinciding with the introduction of post-hoc DL explanation techniques, such as CAM, after 2016. By the end of 2021, there were a total of 60 IIFD-related publications, which then surged to 129 papers over the next two years (2022 and 2023). This ascending trajectory indicates a robust growth in IIFD research, with future projections pointing to an even greater volume of publications in 2024 and beyond. This uptrend is likely driven by continuous advancements in explainable deep learning technologies, signaling a promising direction for further exploration and application in the field.

2) APPLICATION AND CLASSIFICATION

In our comprehensive summary of the applications derived from the literature, we focused on categorizing the findings based on research subjects and specific IIFD tasks. According to Figure 4, bearings emerge as the most frequently studied subject within IIFD research, which also encompasses gearboxes, motors, engines, and a variety of other rotating machinery, including compressors, wind turbines, pump sets,

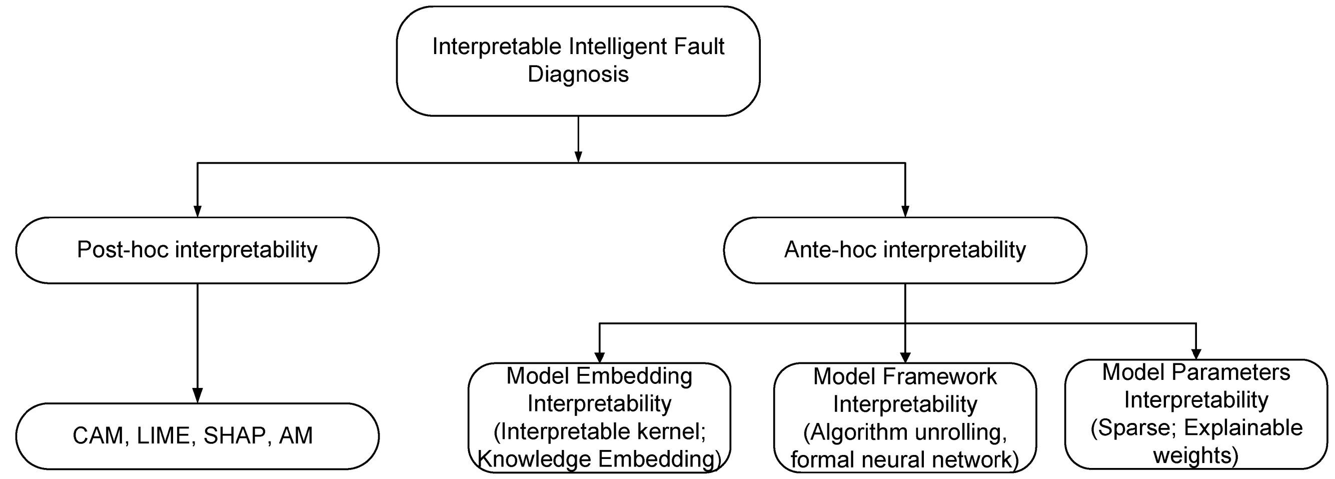

Interpretable Intelligent Fault Diagnosis
Post-hoc interpretability
Ante-hoc interpretability
CAM, LIME, SHAP, AM
Model Embedding Interpretability (Interpretable kernel; Knowledge Embedding)
Model Framework Interpretability (Algorithm unrolling, formal neural network)
Model Parameters Interpretability (Sparse; Explainable weights)

FIGURE 2. The three-level keyword tree.

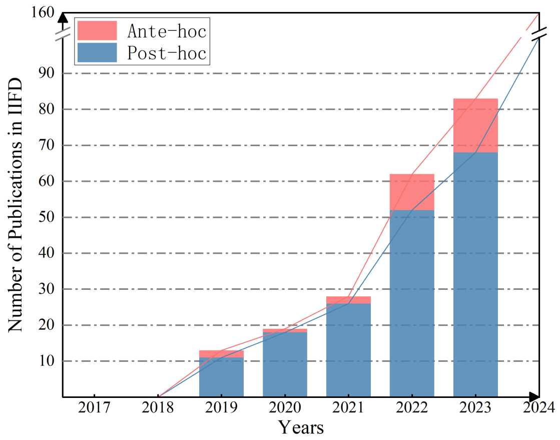


FIGURE 3. Number of publications on ante-hoc and post-hoc IIFD from 2018 to 2023.

and electric motors. Regarding the tasks undertaken in IIFD research, fault classification, and prognostics and health management (PHM) stand out as principal research areas. Among these, bearing fault diagnosis has attracted the most attention, highlighting its critical importance in maintaining the operational integrity of machinery. Conversely, PHM, especially concerning motors and other equipment, has been identified as a promising and valuable research direction. This diversified focus on both research subjects and IIFD tasks underscores the broad applicability and crucial role of IIFD techniques in ensuring the reliability and efficiency of various mechanical systems.

After careful review, this article categorizes the collected papers in relation to the general process of IFD into interpretation based on model frameworks, model structures and mechanisms, and feature importance, as illustrated in Figure 5. Specifically, interpretation based on model frameworks is often implemented during the data preprocessing stage. Integrating traditional signal processing methods with machine learning models can embed specific fault characteristic knowledge and causal links into the diagnostic process. Furthermore, the application of theoretical knowledge to analyze data distributions aids in evaluating model decision boundaries or rules, thereby enhancing interpretability.

Interpretation based on model structures and model mechanisms usually occurs during the feature extraction stage. The model structure interpretability is attained by architecting convolutional kernels that carry physical significance or by unrolling interpretable iterative algorithms into neural networks. Such designs enable extracting features with direct physical relevance, enhancing the ability of the model to make meaningful diagnostic predictions. On the other hand, the interpretability of the mechanism is often achieved through the embedding of explainable mechanisms such as sparsity and logical inference. This approach ensures the model possesses inherent interpretability. These strategies underscore the emphasis on making the feature extraction stage as interpretable as possible, thereby contributing to the overall transparency and effectiveness of IFD models.

Interpretation based on feature importance corresponds to the final feature classification stage. In the process of IFD, the feature classification stage determines the final classification results based on feature importance. However, explanations that focus on feature importance typically rely on passive approaches, such as the use of visualization tools to evaluate if the model has successfully identified features that align with expert knowledge. This passive nature does not actively guide the model towards interpretability. Such reliance on visualization tools for interpretation presents notable limitations, particularly regarding the stability and reliability of the explanation results. Without the ability to influence the interpretative of the model process actively, there is a risk that the explanations provided might not consistently reflect the underlying reasons for the decisions of the model. Consequently, this explanatory approach may raise concerns about the robustness and trustworthiness of the explanation, highlighting the need for more dynamic and interactive explanation mechanisms in the IIFD.

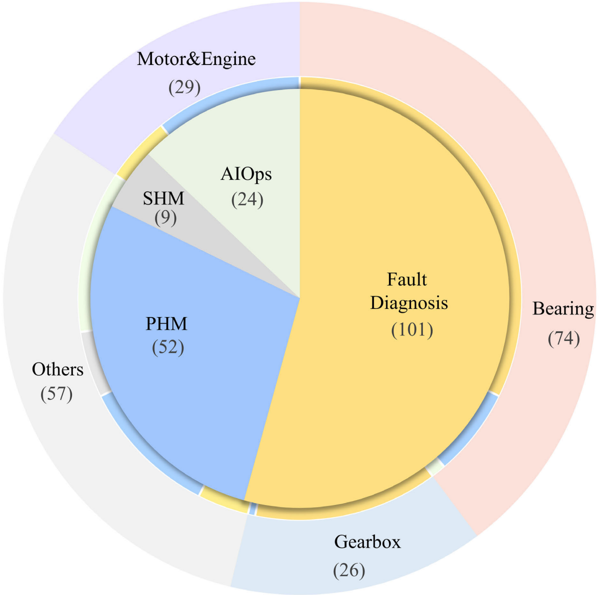


FIGURE 4. Outer ring: research object of the collected literature. Inner circle: main problem distribution solved by the collected literature.

C. COMMONLY USED PUBLIC DATASETS IN IIFD

To achieve reliable and interpretable IFD, access to a substantial amount of high-quality datasets is indispensable. The absence of such data not only hampers the expected performance of intelligent diagnostic models, especially those based on DL, but also undermines the reliability of their interpretive outcomes. The collection of data, predominantly accomplished through various sensors, finds accelerometers for gathering vibration signals as the most widely adopted approach. Nevertheless, procuring sufficient high-quality data in real-world industrial settings poses significant challenges, including a limited number of fault samples, extended collection durations, and elevated costs. In response to these challenges, numerous reputable institutions have made a variety of datasets available for public research and application purposes. This section aims to provide an overview of six datasets commonly utilized in the realm of IIFD for rotating machinery. It is worth noting that current research trends often involve the integration of public datasets with proprietary data to evaluate the efficacy of newly proposed methodologies.

1) CWRU DATASET

The Case Western Reserve University (CWRU) dataset [28], renowned for its application in rotating machinery research, is extensively cited as an open-source benchmark for evaluating various diagnostic methods. As depicted in Figure 6, vibration signals are captured using an accelerometer mounted on the drive end of the motor casing. Data collection encompassed four loading conditions, from 0 to 3 Horsepower (HP), at sampling frequencies of 12 kHz and 48 kHz, with each recording lasting 10 seconds. Faults of four levels (0, 0.007, 0.014, 0.021 inches) and types (rolling element, inner ring, outer ring, healthy) were simulated using electrical discharge machining (EDM).

This dataset enables a 10-class classification task, accommodating studies on variations in loading conditions, fault locations, and severity levels, as outlined in Table 1. While the diagnostic challenges presented by the CWRU dataset are relatively modest—with many models achieving near-perfect accuracy on the 12 kHz signals—this aspect underscores the dataset's quality and thoroughness. However, the high accuracy rates can obscure the comparative analysis of model performances. Despite these considerations, the CWRU dataset's wide application and adaptability render it an indispensable asset for fault diagnosis research, especially for exploring bearing faults across different conditions and failure extents. Access to the CWRU bearing dataset is provided at https://engineering.case.edu/bearingdatacenter.

TABLE 1. Detailed description of CWRU datasets.

Fault Mode	Description
Health State	The normal bearing at 1791rpm and 0HP
Inner ring 1	0.007-inch inner ring fault at 1797rpm and 0HP
Inner ring 2	0.014-inch inner ring fault at 1797rpm and 0HP
Inner ring 3	0.021-inch inner ring fault at 1797rpm and 0HP
Rolling Element 1	0.007-inch rolling element fault at 1797rpm and 0HP
Rolling Element 2	0.014-inch rolling element fault at 1797rpm and 0HP
Rolling Element 3	0.021-inch rolling element fault at 1797rpm and 0HP
Outer ring 1	0.007-inch outer ring fault at 1797rpm and 0HP
Outer ring 2	0.014inch outer ring fault at 1797rpm and 0HP
Outer ring 3	0.021-inch outer ring fault at 1797rpm and 0HP

2) PU DATASET

The Paderborn University (PU) Bearing Data Center offers a comprehensive dataset featuring vibration signals from 32 sets of 6203 Deep Groove Ball Bearings, each with dimensions of 17 × 40 × 12 mm. As detailed in Table 2, these bearings are categorized into three groups based on their condition: six healthy bearings; twelve bearings with artificially induced defects on the inner and outer races through methods such as drilling, EDM, and electric engraving; and fourteen bearings that have incurred natural damage from accelerated lifetime tests.

This dataset is further divided into five categories for detailed analysis: healthy, artificially induced inner ring faults, artificially induced outer ring faults, real inner ring faults, and real outer ring faults. The distribution comprises six healthy bearings, eleven with inner race faults, and twelve

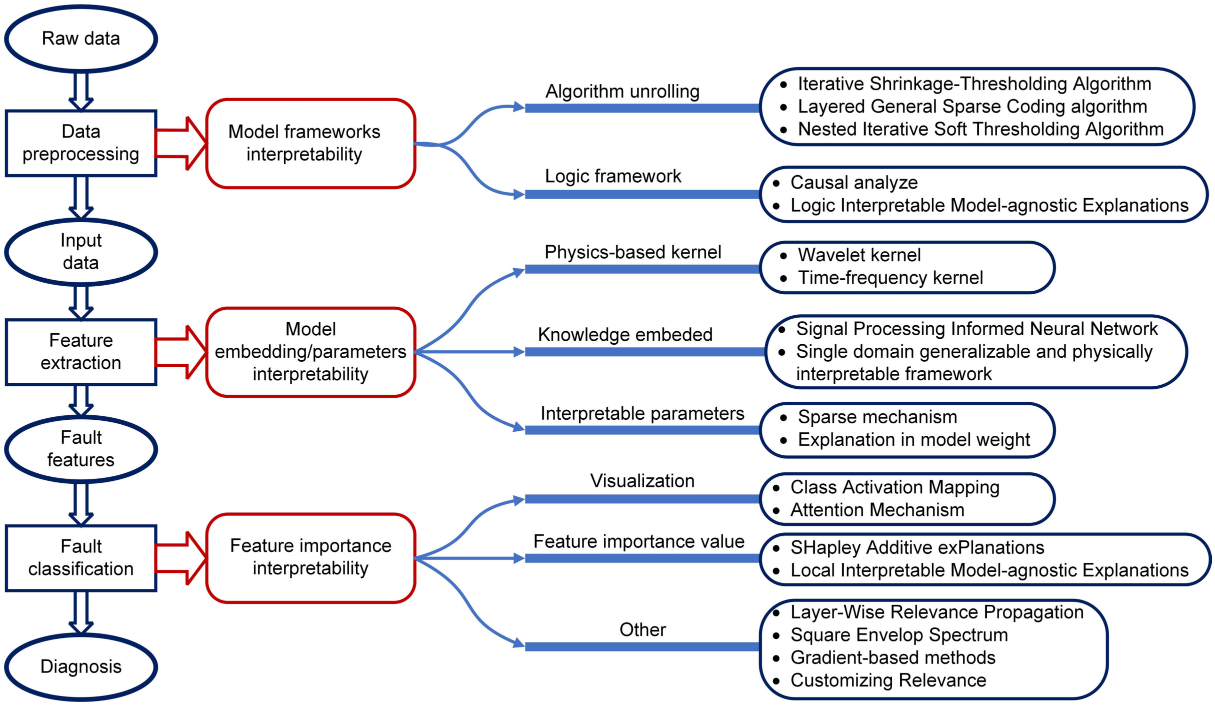

Raw data
Data preprocessing
Input data
Feature extraction
Fault features
Fault classification
Diagnosis
Model frameworks interpretability
Model embedding/parameters interpretability
Feature importance interpretability
Algorithm unrolling
Logic framework
Physics-based kernel
Knowledge embeded
Interpretable parameters
Visualization
Feature importance value
Other
Iterative Shrinkage-Thresholding Algorithm
Layered General Sparse Coding algorithm
Nested Iterative Soft Thresholding Algorithm
Causal analyze
Logic Interpretable Model-agnostic Explanations
Wavelet kernel
Time-frequency kernel
Signal Processing Informed Neural Network
Single domain generalizable and physically interpretable framework
Sparse mechanism
Explanation in model weight
Class Activation Mapping
Attention Mechanism
SHapley Additive exPlanations
Local Interpretable Model-agnostic Explanations
Layer-Wise Relevance Propagation
Square Envelop Spectrum
Gradient-based methods
Customizing Relevance

FIGURE 5. The classification of collected literature along with the process of IFD.

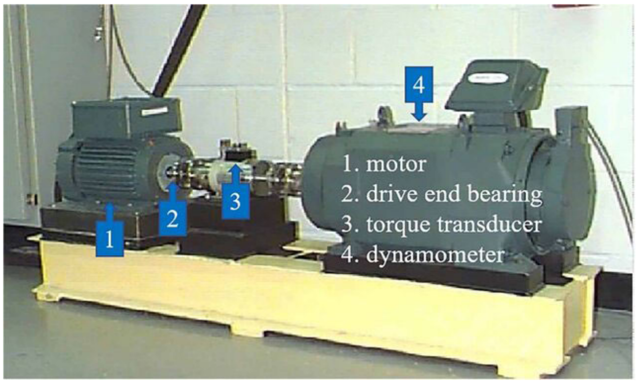


FIGURE 6. CWRU bearing experimental system.

with outer race faults. Bearings presenting both inner and outer race faults—totaling three—are excluded from this count due to the complexity of their damage.

In their research, Paderborn University investigators categorized these bearings based on the predominant fault. For instance, a bearing is classified under inner race faults if the damage to the inner race exceeds that to the outer race. This dataset is instrumental for conducting comparative studies and for research on information fusion from different sensor signals, offering a unique opportunity to compare real-world faults against artificial ones. The PU-bearing dataset is accessible for download at https://mb.unipaderborn.de/kat/forschung/datacenter/bearing-datacenter/.

3) MFPT BEARING DATASET

The MFPT dataset, made available by the Society for Machinery Failure Prevention Technology [29], encompasses three experimental sets of bearing vibration data: a baseline dataset, an inner race fault set, and an outer race fault set. The baseline dataset comprises three files, each sampled at 97,656 Hz over a duration of 6 seconds under a 270-pound load. Both the inner and outer race fault sets consist of seven files, captured at 48,828 Hz for 3 seconds across seven distinct load conditions: 0 lbs (for the inner race) / 25 lbs (for the outer race), 50 lbs, 100 lbs, 150 lbs, 200 lbs, 250 lbs, and 300 lbs. These measurements were recorded using a single-channel radial accelerometer.

The bearings used in the MFPT experiments are deep groove ball bearings, characterized by a pitch diameter of 31.62 mm, a ball diameter of 5.97 mm, a contact angle of 0°, and an element number of 8. The collected signals were segmented into 50 samples for each load condition, yielding a total of 350 samples across the seven load types. Each sample contains 2,000 data points. Subsequently, the dataset is categorized into three classes. As detailed in Table 3, bearings are classified into one healthy state and two fault states—inner ring fault and rolling element fault—across 15 categories (one healthy state and 14 fault states) based on the load conditions. This dataset is invaluable for examining the dynamic behavior under varying operational conditions. The MFPT-bearing dataset is accessible for download at https://www.mfpt.org/fault-data-sets/.

4) IMS DATASET

The IMS bearing datasets, produced by the NSF I/UCR Center for Intelligent Maintenance Systems (IMS) [30], are derived from three test-to-failure experiments. These datasets meticulously document four distinct bearing fault

TABLE 2. Detailed description of PU datasets (S: single damage; R: repetitive damage; M: multiple damage).

Bearing Code	Fault Mode	Description
K001	Health state	Run-in 50h before test
K002	Health state	Run-in 19h before test
K003	Health state	Run-in 1h before test
K004	Health state	Run-in 5h before test
K005	Health state	Run-in 10h before test
K006	Health state	Run-in 16h before test
KA01	Artificial outer ring fault (Level 1)	Made by EDM
KA03	Artificial outer ring fault (Level 2)	Made by electric engraver
KA05	Artificial outer ring fault (Level 1)	Made by electric engraver
KA06	Artificial outer ring fault (Level 2)	Made by electric engraver
KA07	Artificial outer ring fault (Level 1)	Made by drilling
KA08	Artificial outer ring fault (Level 2)	Made by drilling
KA09	Artificial outer ring fault (Level 2)	Made by drilling
KI01	Artificial inner ring fault (Level 1)	Made by EDM
KI03	Artificial inner ring fault (Level 1)	Made by electric engraver
KI05	Artificial inner ring fault (Level 1)	Made by electric engraver
KI07	Artificial inner ring fault (Level 2)	Made by electric engraver
KI08	Artificial inner ring fault (Level 2)	Made by electric engraver
KA04	Outer ring damage (single point + S + Level 1)	Caused by fatigue and pitting
KA15	Outer ring damage (single point + S + Level 1)	Caused by plastic deform and indentation
KA16	Outer ring damage (single point + R + Level 2)	Caused by fatigue and pitting
KA22	Outer ring damage (single point + S + Level 1)	Caused by fatigue and pitting
KA30	Outer ring damage (distributed + R + Level 1)	Caused by plastic deform and indentation
KB23	Outer ring and inner ring damage (distributed + M + Level 3)	Caused by fatigue and pitting
KB24	Outer ring and inner ring damage (distributed + M + Level 3)	Caused by fatigue and pitting
KB27	Outer ring and inner ring damage (distributed + M + Level 1)	Caused by plastic deform and indentation
KI04	Inner ring damage (single point + M + Level 1)	Caused by fatigue and pitting
KI14	Inner ring damage (single point + M + Level 1)	Caused by fatigue and pitting
KI17	Inner ring damage (single point + R + Level 1)	Caused by fatigue and pitting
KI18	Inner ring damage (single point + S + Level 2)	Caused by fatigue and pitting
KI21	Inner ring damage (single point + S + Level 2)	Caused by fatigue and pitting

TABLE 3. Detailed description of MFPT datasets.

Label	Health State	Load	Fault Mode	Description
0	Health	270lbs	Health Sate	Fault-free bearing working at 270lbs
1	Outer race	25lbs	Outer ring 1	Outer ring fault bearing working at 25lbs
2	Outer race	50lbs	Outer ring 2	Outer ring fault bearing working at 50lbs
3	Outer race	100lbs	Outer ring 3	Outer ring fault bearing working at 100lbs
4	Outer race	150lbs	Outer ring 4	Outer ring fault bearing working at 150lbs
5	Outer race	200lbs	Outer ring 5	Outer ring fault bearing working at 200lbs
6	Outer race	250lbs	Outer ring 6	Outer ring fault bearing working at 250lbs
7	Outer race	300lbs	Outer ring 7	Outer ring fault bearing working at 300lbs
8	Inner race	0lbs	Outer ring 1	Inner ring fault bearing working at 0lbs
9	Inner race	50lbs	Inner ring 2	Inner ring fault bearing working at 50lbs
10	Inner race	100lbs	Inner ring 3	Inner ring fault bearing working at 100lbs
11	Inner race	150lbs	Inner ring 4	Inner ring fault bearing working at 150lbs
12	Inner race	200lbs	Inner ring 5	Inner ring fault bearing working at 200lbs
13	Inner race	250lbs	Inner ring 6	Inner ring fault bearing working at 250lbs
14	Inner race	300lbs	Inner ring 7	Inner ring fault bearing working at 300lbs

conditions: rolling element fault, inner ring fault, outer ring fault, and healthy status. The bearings, installed at various locations, all failed beyond their anticipated service life. Data was captured at a 20 kHz sampling frequency, with each recording lasting 1 second, resulting in a collection ranging from 984 to 4,448 sample files. It's critical to note that these failures should not be oversimplified into merely three categories due to the variability in fault locations within the bearings [31]. For research and analysis purposes, the IMS bearing dataset is accessible for download at https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/.

5) XJTU-SY DATASET

The XJTU-SY bearing datasets, a collaborative effort between Xi'an Jiaotong University and Changxing Sumyoung Technology Company, encompass data from fifteen run-to-failure experiments on bearings under three different operational conditions. Each dataset contains 32,768 data points for every bearing, with a sampling frequency of 25.6 kHz, amounting to 1.28 seconds of vibration data collected over one minute for detailed analysis. These datasets are organized into fifteen classes, each corresponding to a specific fault diagnosis, enabling comprehensive investigations into bearing performance and failure dynamics. The selection of data aims to capture the end-of-life phase of the bearings, enhancing the study of failure modes and diagnostic techniques.

Comprehensive details about the operational lifespan and failure types of each bearing are meticulously documented in Table 4. Moreover, the fault characteristic frequency (FCF) of each bearing is calculated and included in Table 4, reflecting the test conditions and parameters. The experimental setup is depicted in Figure 7. For research and analysis purposes, the XJTU-SY datasets are available for download at https://biaowang.tech/xjtusy-bearing-datasets/.

TABLE 4. Detailed description of XJTU-SY datasets.

Condition	File	Lifetime	Fault element	FCF(Hz)
Speed: 35 Hz Load: 12kN	Bearing1_1	2h 3min	Outer ring	107.9
	Bearing1_2	2h 41min	Outer ring	107.9
	Bearing1_3	2h 38min	Outer ring	107.9
	Bearing1_4	2h 2min	Cage	13.5
	Bearing1_5	5h 2min	Inner ring and Outer ring	172.1, 107.9
Speed: 37.5 Hz Load: 11kN	Bearing2_1	8h 11min	Inner ring	184.4
	Bearing2_2	2h 41min	Outer ring	115.6
	Bearing2_3	8h 53min	Cage	14.5
	Bearing2_4	42min	Outer ring	115.6
	Bearing2_5	5h 39min	Outer ring	115.6
Speed: 40 Hz Load: 10kN	Bearing3_1	42h 18min	Outer ring	123.3
	Bearing3_2	41h 36min	Inner and Outer ring, Rolling element and Cage	15.4, 82.7, 196.7, 123.3
	Bearing3_3	6h 11min	Inner ring	196.7
	Bearing3_4	25h 15min	Inner ring	196.7
	Bearing3_5	1h 54min	Outer ring	123.3

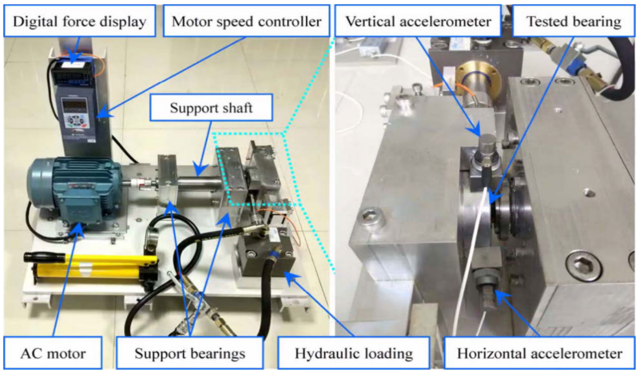


FIGURE 7. Test rig that produces the XJTU-SY bearing dataset.

6) SEU DATASET

The Southeast University (SEU) dataset, as detailed in the study by Shao et al. [32], comprises both bearing and gear datasets collected using a Drivetrain Dynamic Simulator. This dataset is meticulously structured to replicate two distinct operational scenarios defined by a Rotating Speed - Load Configuration: namely, 20 Hz-0 V and 30 Hz-2 V. Within this dataset, gear conditions are divided into five categories: healthy, chipped tooth, missing tooth, root fault, and surface fault. Similarly, bearing conditions are also classified into five categories: healthy, inner ring fault, outer ring fault, combined inner and outer ring faults, and rolling element fault. Collectively, the SEU dataset delineates 20 distinct health states, with comprehensive details provided in Table 5.

This dataset is invaluable for conducting time-frequency domain analyses of operational conditions and diagnosing faults in mechanical components. Each file in the SEU dataset comprises seven columns of vibration signals and a single column of motor torque signals, thus providing a rich dataset for thorough analysis. The SEU gearbox dataset is publicly available and can be accessed for research via the following GitHub repository link: https://github.com/cathysiyu/Mechanical-datasets.

TABLE 5. Detailed description of SEU datasets.

Fault Mode	RS-LC	Fault Mode	RS-LC
Health Gear	20Hz - 0V	Health Bearing	20Hz - 0V
Health Gear	30Hz - 2V	Health Bearing	30Hz - 2V
Chipped Tooth	20Hz - 0V	Inner ring	20Hz - 0V
Chipped Tooth	30Hz - 2V	Inner ring	30Hz - 2V
Missing Tooth	20Hz - 0V	Outer ring	20Hz - 0V
Missing Tooth	30Hz - 2V	Outer ring	30Hz - 2V
Root Fault	20Hz - 0V	Inner + Outer ring	20Hz - 0V
Root Fault	30Hz - 2V	Inner + Outer ring	30Hz - 2V
Surface Fault	20Hz - 0V	Rolling Element	20Hz - 0V
Surface Fault	30Hz - 2V	Rolling Element	30Hz - 2V

7) SUMMARY

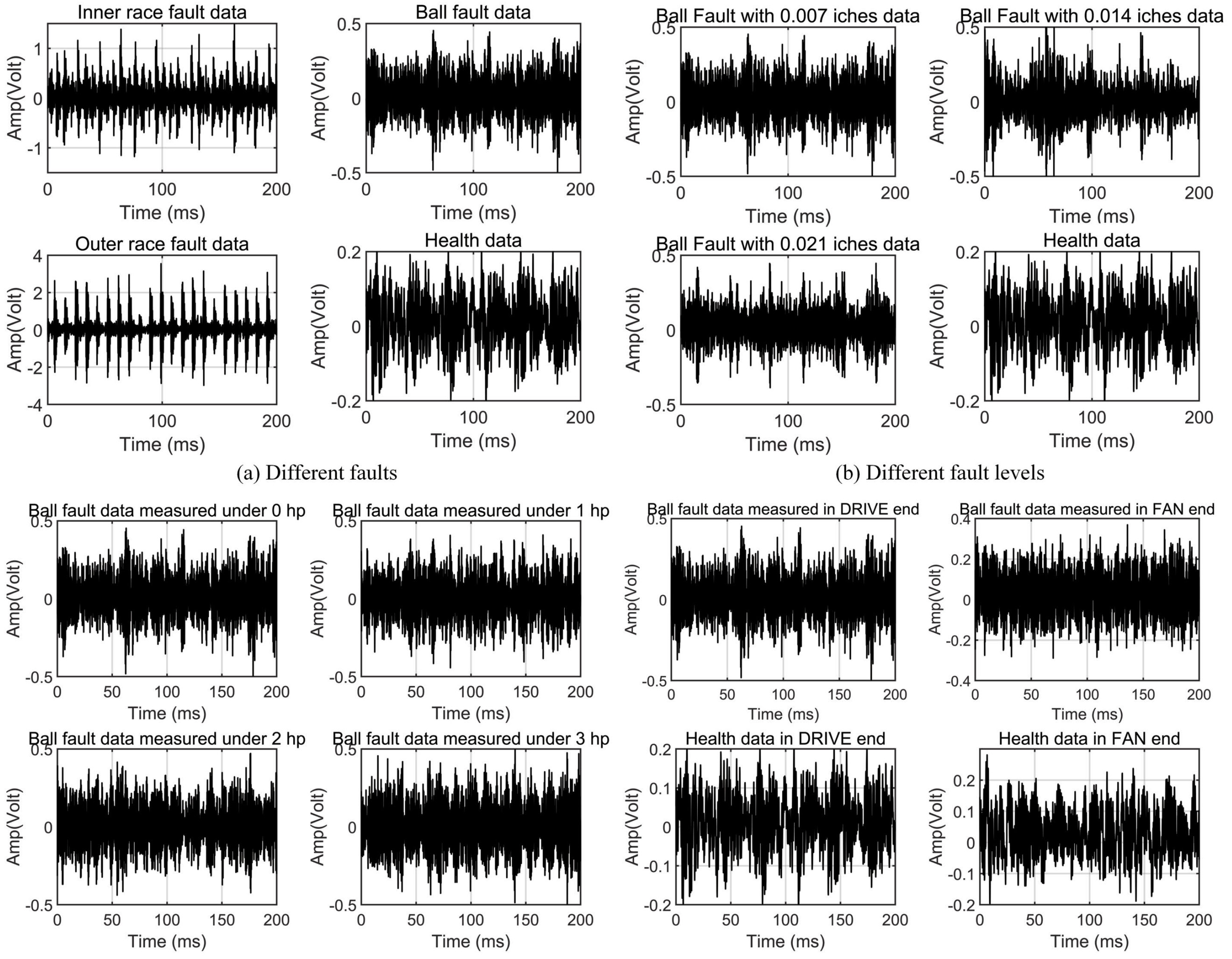
This paper provides a selective overview of datasets pertinent to rotating machinery vibration signals, focusing on those that are broadly recognized within the IFD community. A notable example, the CWRU dataset, enjoys particular esteem among researchers. Figure 8 visually contrasts segments of vibration signals from the CWRU dataset under various operational scenarios. Such comparisons illuminate the substantial effects that fault type, severity, operational conditions, and sensor positioning have on the gathered data. These factors, in turn, significantly influence the performance of fault diagnosis models. Analogous to the CWRU, other datasets in the field present distinct features that can shape the results of fault diagnosis studies. The intent of this section is to guide researchers towards datasets that best match their specific research objectives, thereby enabling more effective contributions to the fault diagnosis discipline. Notably, Zhao et al. [31] conducted an open source benchmark study on these datasets and provided the tutorial code via the following GitHub repository link: https://github.com/ZhaoZhibin/DL-based-Intelligent-Diagnosis-Benchmark.

III. POST HOC INTERPRETATION OF IFD

This section details the post-hoc interpretability methods used in IIFD. It offers insights into their application areas and outlines the fundamental principles that guide these methods, enhancing understanding.

A. OVERVIEW

Contemporary fault diagnosis techniques predominantly leverage neural networks and other advanced methods for



(a) Different faults

(b) Different fault levels

(c) Same fault under different load conditions

(d) Same fault measured at different position

FIGURE 8. The waveforms of vibration signals from the CWRU dataset. (a) Different faults (b) Different fault levels (c) Same fault under different loading conditions (d) Same fault measured at different positions.

fault classification. Yet, these approaches frequently rely on "black-box" models, characterized by their lack of transparency and inability to elucidate the reasoning behind their classifications. To address this issue, researchers have turned to XAI strategies to interpret classification outcomes or identify underlying patterns. Such interpretive efforts, facilitated by XAI, are commonly classified under the umbrella of post-hoc interpretability.

Presently, notable post-hoc interpretability methods in fault diagnosis include CAM, SHAP, LIME, among others. The following subsections will delve into these methodologies in greater detail.

B. CAM-BASED POST HOC INTERPRETATION

CAM is a technique designed to elucidate neural network decisions through the visualization of learned features, highlighting their influence—be it positive or negative—on predictions [28]. In this subsection, we succinctly examine CAM's theoretical underpinnings and its application within the realm of IIFD.

1) A BRIEF INTRODUCTION TO CAM

As shown in Figure 9, CAM is generated by calculating the weighted sum of the feature maps from the final convolutional layer [33]. The weights are between the global average pooling(GAP) and output, which computes the spatial average for each unit in the feature map of the last convolutional layer.

According to Ref. [33], CAM could be described as (1).

$$M_C(x, y) = \sum_k w_k^c f_k(x, y). \tag{1}$$

where $M_C(x, y)$ is the result of CAM in category c, w_k^c is the weight of the kth feature map, and $f_k(x, y)$ is the kth convolved feature map at position (x, y).

CAM is designed explicitly for CNN architectures that apply GAP to convolutional maps before prediction. To extend CAM's applicability to broader network types, Selvaraju et al. [34] developed Gradient-weighted CAM (Grad-CAM). This method utilizes the gradient information flowing between the last convolutional layer and the output nodes to gauge the significance of each neuron in making a specific decision. The operation of Grad-CAM is detailed in equations (2) and (3).

$$\alpha_q^c = \frac{1}{z}\sum_i \sum_j \frac{\partial y^c}{\partial A_k}. \tag{2}$$

$$L_{Grad-CAM}^c = RELU(\sum_k \alpha_k^c A_k) \tag{3}$$

Instead of GAP, Grad-CAM used α_q^c as the importance weight of the k feature graphs. α_q^c is obtained by averaging the gradients of the k feature graphs. Finally, the obtained $L_{Grad-CAM}^c$ is resampled to match the input sample size.

In recent years, many visualization methods based on CAM or Grad-CAM have been proposed, such as Grad-CAM++ [35], Eigen-CAM [36], and others.

2) APPLICATIONS OF CAM TO IIFD

The utilization of CAM and its derivatives in IIFD is meticulously cataloged in Table 6. Specifically, Grad-CAM emerges as a cornerstone post-hoc interpretability technique within the realm of machinery fault diagnosis. For instance, Brito et al. [37] successfully applied Grad-CAM to elucidate the workings of a fault diagnosis model for rotating machinery, which was notably enhanced with synthetically augmented data. This demonstrated the method's practicality in clarifying model decisions. Similarly, Lin and Jhang [38] merged Grad-CAM with original signal data to produce attention maps, offering deep insights into the rationale behind specific bearing fault model classifications.

Addressing a common challenge in traditional fault diagnosis, where identifying defect frequencies often requires deep domain knowledge. Yoo and Jeong [39] innovatively applied Grad-CAM. This approach allowed for the visualization of CNN activation regions to determine defect frequencies directly, circumventing the need for expert input and also pinpointing the most effective operational layer for Grad-CAM application. Expanding the scope of model interpretability, Lu et al. [40] employed Grad-CAM to evaluate the significance of training samples, thus broadening interpretability to include not just feature importance but also the relevance of specific training samples in model training. Chen and Lee [41] utilized Grad-CAM to generate heat maps of the time-frequency domain features of CNN. They validated the interpretative results of Grad-CAM through NN, adaptive network-based fuzzy inference system (ANFIS), and Decision-Tree and discovered machine learning pays more attention to high-frequency features. Saeki et al. [42] employed Grad-CAM to interpret results from a CNN-based anomaly detection system for rotating machinery, evaluating the Grad-CAM by comparing them with the expert diagnostic.

Beyond these applications, the field has seen the development and adoption of advanced CAM variations like Grad-CAM++, Score-CAM, and Eigen-CAM to further enhance interpretability in IIFD. Chen et al. [16] embedded 1-D Grad-CAM++ in the model to identify regions of interest in the convolutional layer, combining prior knowledge of bearing faults to comprehend the learned features and model decisions. Lan et al. [43] applied Grad-CAM++ to illustrate the interpretability of the proposed model by visualizing the saliency map. Yu et al. [44] employed Eigen-CAM to provide intuitive explanations for the fault diagnosis results of ResNets, demonstrating the capacity of the model to accurately capture fault and Eigen-CAM outperforming Grad-CAM. In order to improve the performance of the CAM-based method in fault diagnosis, researchers introduced the optimization architecture for CAM. Yang et al. [45] designed a located loss in CNN to drive the model to learn primary features. They explained that the model decisions originate from these primary features through SS-CAM. Li et al. [28] proposed the Multilayer Grad-CAM, which can effectively extract periodic pulses in time-domain signals. Simultaneously, it clearly displays different bearing fault characteristic frequencies in the spectrum, addressing the issue of decreasing feature resolution of Grad-CAM with deeper networks on vibration signals. Additionally, they defined three metrics (RATM, RATA, CEI) to quantify the interpretability of deep neural networks. Chen et al. [46] designed GS-CAM which combine the Grad-CAM and Score-CAM to analyze the attention distribution of the proposed model on time-domain signals. Kim et al. [47] propose the frequency-domain-based grad-CAM to visualize the classification criteria in the frequency domain using the learned network in the time domain.

Despite these advancements, CAM-based approaches exhibit limitations in their adaptation to regression tasks. A significant limitation of CAM-based approaches is their potential inability to fully capture temporal aspects of data, which is critical for fault diagnosis involving sequential or time-series analysis. Additionally, gradient-based CAM methods also have additional drawbacks. On the one hand, the weights obtained from the gradient-based CAM can not provide the right confidence scores for the feature maps, leading to coarse localization saliency maps with Grad-CAM when the input data contains numerous essential features. On the other hand, gradient-based CAM for CNNs may inadvertently focus on irrelevant parts of the data due to gradient saturation in the flat zero-gradient regions of the ReLU activation function. These limitations highlight the complexities and potential areas for improvement in employing CAM-based methods for fault diagnosis, especially in accurately interpreting and localizing fault-relevant features in complex data scenarios.

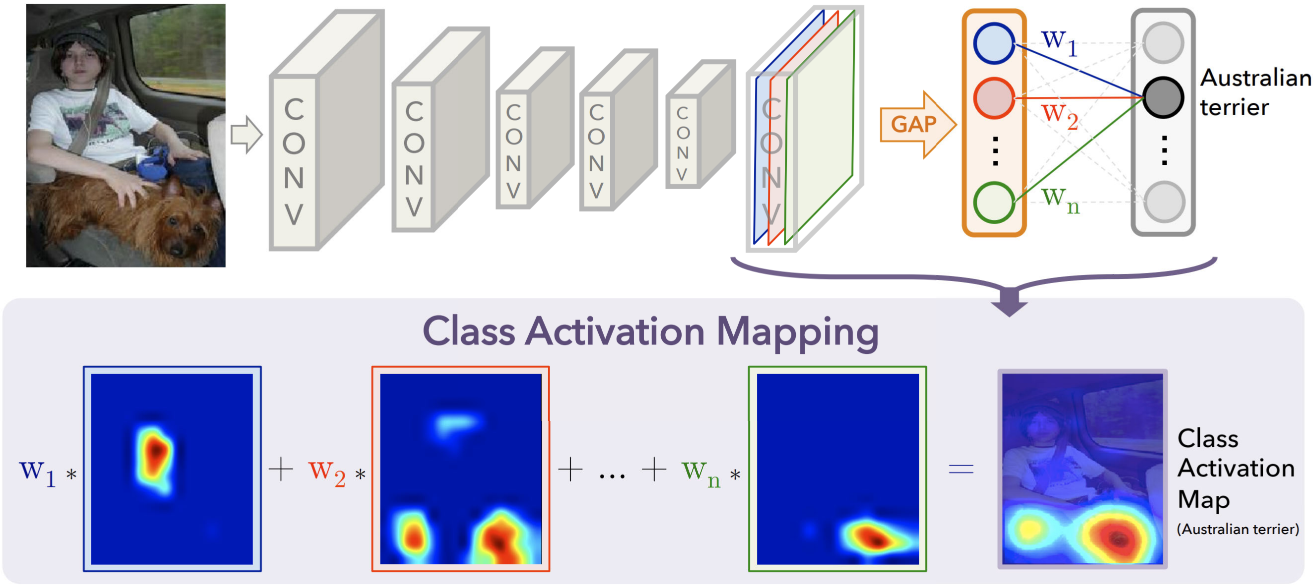


FIGURE 9. The principle of Class Activation Mapping:the predicted class score is mapped back to the previous convolutional layer to generate the CAMs. The CAM highlights the class-specific discriminative regions [33].

TABLE 6. The application of CAM and customed CAM to IIFD.

Object	Methodologies	Ref.
Bearing	Popular CAM	Hu [48], Brito [37], Han [49], Chen [41], Lu [40], Saeki [42], Yu [44], Lin [38], Feng [50], Chen [16], He [51], Zhang [20], Wen [52], Guo et al. [53]
	Customed-CAM	Chen [46], Sun [54], Kim [47], Yang [45], Li [28]
Gearbox	Popular CAM	Hu [48], Brito [37], Hao [55], Lan [43], Zhang [20], Liu [56]
	Customed-CAM	Chen [46], Yang [45]
Motor	Popular CAM	Oh [57] , Yoo [39], Mey [58]
	Customed-CAM	Kim [47]
Others	Popular CAM	Huh [59], Han [49], Liu [60], Liu [61], Kim [62], Ren [63], Li [64], Chao [65], Cheng [66], Dopierała [67], Nasiri [68], Lee [69], Li [70], Ardito [71]
	Customed-CAM	Liu [72], Li et al. [73]

C. LIME-BASED POST HOC INTERPRETATION

LIME is a technique that explains the predictions of any machine learning model by approximating it locally with an interpretable model, thus allowing for the understanding of individual predictions regardless of the original model's complexity. This technique involves creating a simple, interpretable model (such as a linear model or decision tree) that approximates the behavior of the complex model within the vicinity of the instance being explained. Therefore, LIME provides insight into which features were most influential for a particular prediction, enhancing the transparency and trustworthiness of the model on a case-by-case basis. We briefly review LIME and summarize its applications in IIFD.

1) A BRIEF INTRODUCTION TO LIME

LIME facilitate the identification of an interpretable model, $g \in G$, within a local scope based on an interpretable representation [74]. Here, G represents the collection of potential interpretable models, while $\Omega(g)$ denotes the complexity of model g. As illustrated in Figure 10, for a given sample x evaluated by g, with f being the actual model and $f(x)$ the predicted probability by g, LIME introduces perturbations in the vicinity of x to generate a perturbed sample z. The proximity measure $\pi_x(z)$ quantifies the distance between x and its perturbation z. The objective is to minimize the loss function $L(f, g, \pi_x)$, which measures the discrepancy between f and g within the vicinity of x, while also considering the complexity $\Omega(g)$, as shown in:

$$\pi_x(z) = \arg\min L(f, g, \pi_x) + \Omega(g). \tag{4}$$

Given the challenges in computing this in the original image dimension, LIME employs Super Pixels to transform x into a binary representation $x' \in \{0, 1\}^{d'}$, subsequently generating z'. The proximity measure $\pi_x(z)$ is defined as:

$$\xi(x) = \exp\left(-\frac{D(x, z)^2}{\sigma^2}\right). \tag{5}$$

Here, $D(x, z)$ represents the distance between x and z, and σ is the kernel width of the perturbation. With (5), after reverting z' back to the original dimension, computing $f(z)$, and evaluating $\pi(x)$, the loss function L in (4) can be

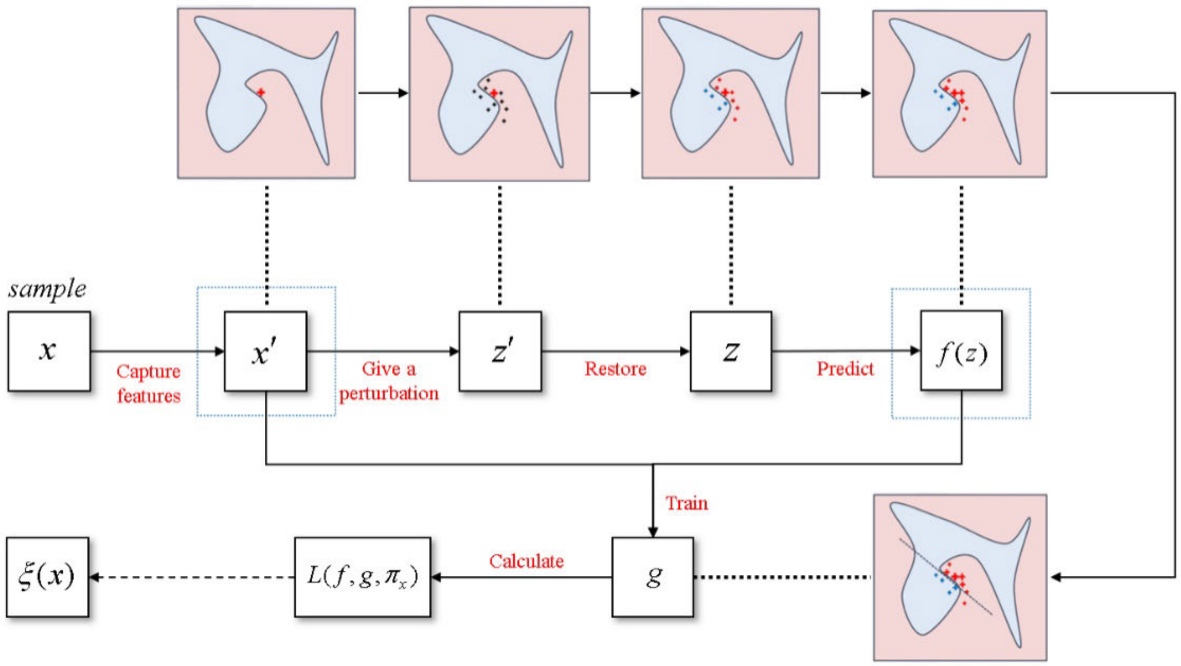


FIGURE 10. The principle of LIME: an algorithm used for explaining predictions of machine learning models. The process starts with a sample data point, x, which goes through a feature extraction phase to produce x'. The algorithm then perturbs x' to generate z', a version of the data with slight variations. The perturbed data z' is then restored back to its original representation, z, after which the machine learning model makes a prediction $f(z)$. Concurrently, LIME calculates the explanation model $\xi(x)$ using the loss function $L(f, g, \pi_x)$, which considers the fidelity of the local model g to the predictions of the global model f, weighted by the proximity π_x of the perturbed samples to the original sample x. The explanation model g is trained to approximate $f(z)$ locally around x, thereby providing interpretable insights into the model's prediction for x.

reformulated as:

$$L(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(f(z) - g(z'))^2. \tag{6}$$

The specific implementation of $g(z')$ depends on the chosen class G, for instance, in linear systems, $g(z') = \omega_g \cdot z'$, where ω_g is the weight derived from distance. Through the optimization of (4), LIME can train a local interpretable model based on the interpretable feature space, offering a pragmatic approach to understanding model predictions on a granular level.

2) APPLICATION OF LIME TO IIFD

As a post-hoc interpretability method, LIME has seen extensive application in IIFD, with relevant studies summarized in Table 7. This research is categorized based on data collection methods and diagnostic targets, primarily focusing on bearings, gears, motors, and engines. Unlike CAM, which visualizes feature importance across the model, LIME delves into explaining individual predictions, shedding light on the features crucial for specific decision-making instances. In the realm of IIFD: Lu et al. [11] developed a 1-D CNN model combined with LIME for accurate fault classification in rolling bearings under various speeds. Wang et al. [75] created a predictive model for evaluating the rolling contact fatigue in martensitic steel, utilizing LIME to ensure the model's interpretability and reliability. Several enhancements to LIME have been proposed to augment its interpretability. Recio-García et al. [76] improved the data generation process with a case-based reasoning method. Saini and Prasad [77] optimized LIME's sampling strategy using the Gaussian process. Zafar and Khan [78] applied Hierarchical Clustering and K-Nearest Neighbor for data grouping instead of random perturbation. Dikopoulou et al. [79] introduced a graphical methodology for achieving global model-agnostic interpretability with LIME. Xiang et al. [80] enhanced LIME's stability and local fidelity through a variational autoencoder.

LIME is prized for its local fidelity, simplicity, and broad compatibility, offering a nuanced understanding of black-box models. However, its application in IIFD faces challenges. Firstly, the selection of samples in LIME may require expert judgment, introducing potential biases. Secondly, LIME cannot fully represent the original model, with its effectiveness heavily model-dependent. Lastly, the methodology for determining weights and the extensive computation required for each model analysis [81] pose significant limitations. Consequently, despite its theoretical appeal, LIME's practical deployment in IIFD has been more theoretical than empirical in recent years.

D. SHAP-BASED POST HOC INTERPRETATION

SHAP is a method that assigns each feature an importance value for a particular prediction, based on the concept

TABLE 7. Summary of LIME used in IIFD.

Objects	References	Data collection approach
Bearings	Alfeo et al. [82]	DMF
	Lu et al. [11], Sanakkayala et al. [83], Pham et al. [84]	DIF
Gears	Amin et al. [85], [86], Mai et al. [87], Wang et al. [75]	DIF
Motors	Yao et al. [88], Srinivasan et al. [89], Akpudo et al. [90], Yu et al. [91]	DMF
	Kim et al. [92]	DIF
Engine	Yang et al. [93], Protopapadakis et al. [94], Udo Sass et al. [95], Kobayashi et al. [96], Serradilla et al. [97], Baptista et al. [98]	DMF
Others	Al-Zeyadi et al. [99], Khan et al. [100], Widianto et al. [101], Usuga-Cadavid et al. [102], Madhikermi et al. [103], Sharma et al. [104], Liang et al. [105], Sairam et al. [106], Ferraro et al. [107]	DMF
	Pandey et al. [108], Tang et al. [109], Zhang et al. [110], Hanchate et al. [111], Onchis et al. [112], Sairam et al [113]	DIF

DMF:Data Collected from Multiple Features,
DIF:Data Collected from Identical Features

of Shapley values from cooperative game theory, thereby offering a consistent and locally accurate interpretation of the model's output. The flow chart of SHAP is shown in Figure 11. This section provides a concise overview of SHAP and discusses its applications within the context of IIFD.

1) A BRIEF INTRODUCTION TO SHAP

SHAP, introduced by Lundberg et al. [114], offers a refined methodology for interpreting predictions from machine learning models. It works by analyzing a trained model's input and output, attributing specific contributions—known as SHAP values—to each input feature based on their impact on the model's prediction. This process involves tracing the prediction back to its input features in a layer-wise manner.

The computation of SHAP values is defined by the following equation:

$$\Phi_i = \sum_{S \subseteq F/i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_S(x) - f_{S/i}(x)]. \quad (7)$$

In this equation, Φ_i represents the SHAP value for feature i, indicating its relative contribution to the prediction. SHAP utilizes various approximation methods to accommodate different machine learning architectures, including Kernel SHAP for general models, Tree SHAP for tree-based models, and Deep SHAP for deep learning networks. Each approach, grounded in the principle expressed in (7), aims to quantify the influence of individual features on the prediction.

1) Kernel SHAP offers a universal solution applicable to a broad range of models.
2) Tree SHAP provides specialized analysis for tree-based models, enhancing efficiency and accuracy.
3) Deep SHAP tailors its analysis for deep learning models, adapting to their complexity.

The choice among these SHAP variants depends on the model's structure and the specific interpretability requirements. Together, these methods strive to make model predictions transparent, thereby allowing SHAP to illuminate complex model behaviors with a solid mathematical foundation.

2) APPLICATIONS OF SHAP TO IIFD

SHAP has been an effective way in the research of post-hoc IIFD. Asutkar and Tallur [115] enhanced the fault detection strategy using SHAP to identify the most prominent features contributing to fault detection. Yao et al. [116], [117], [118], [119] classified different types of faults based on ensemble methods (such as Random Forest (RF), Gradient Boosting, and AdaBoost) and used SHAP to explain the classification results of the model. Kumar and Hati [120] proposed a deep CNN model based on an adaptive gradient optimizer and conducted a SHAP analysis to interpret the vibration images and decision-making process of the proposed model. Hasan et al. [121] introduced a data preprocessing method utilizing the Stockwell Transformation Coefficient, followed by an interpretable feature selection using RF. The fault diagnosis was then conducted using a K-NN classifier, with the diagnostic results of K-NN being interpreted according to SHAP. Brito et al. [122] utilized SHAP to prioritize the importance of features, providing interpretation and analysis of the results derived from the unsupervised anomaly detection model.

To better explain deep learning models, researchers have made improvements to SHAP. Yao et al. [123] used the integration of SHAP with DeepLIFT in the form of Deep-SHAP, an algorithm better suited for deep learning models characterized by high non-linearity and complex layer structures, and has successfully automated the extraction of fault characteristic frequencies. Wang and Wang [124] used SHAP to interpret the results of motor fault diagnosis with SVM, RF, and NN and found that the average vibration frequency is the most critical feature in diagnosing motor faults. In addition, SHAP is also used for process control [125], machine predictive maintenance [107], [126], [127], [128], [129], and sensor fault diagnosis [64], [130], [131], among other applications.

Compared to LIME, which constructs local linear approximations to explain individual predictions, SHAP elucidates the decision-making behavior of single predictions and interprets the significance of features in the entire fault diagnosis model. This approach offers a more comprehensive and precise overall model explanation [132], providing deeper insights into the model's workings. However, it is

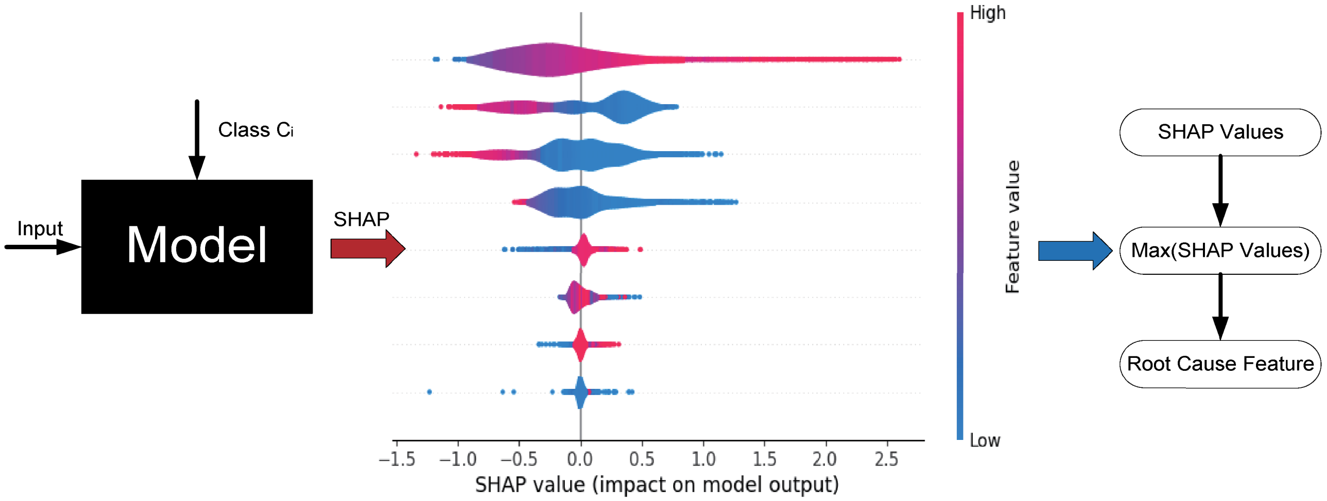


FIGURE 11. The flow chart of SHAP: Inputs produce predictions for class C_i, and SHAP values determine each feature's influence on that prediction. The chart shows the SHAP value distribution across features, with the magnitude indicating impact strength and color representing feature value. The process isolates the "Root Cause Feature" by identifying the feature with the highest SHAP value, highlighting the most influential factor in the model's output.

TABLE 8. The application of SHAP and customed SHAP to IIFD.

Object	Methodologies	Ref.
Bearing	Kernel SHAP	Asutkar et al. [115], Pham et al. [116], Sinha et al. [117], Brusa et al. [118], Daiki et al. [119], Kumar et al. [120], Hasan et al. [121], Brito et al. [122], An et al. [133], Bindingsbø [134] Yao et al. [123]
	Extension	Hu et al. [135]
Gear	Kernel SHAP	Herwig et al. [136], Zhang et al. [137]
	Extension	Hu et al. [135]
Motor & Engine	Kernel SHAP	Movsessian et al. [138], Kirchgassner et al. [127], Sasaki et al. [139], Vitor et al. [140], Pan et al. [141], Youness et al. [142], Baptista et al. [143]
	Extension	Wang et al. [124]
Others	Kernel SHAP	Szelążek et al. [144], Steurtewagen et al. [126], Gu et al. [128], Ramezani et al. [129], Hwang et al. [130], Fang et al. [131], Li et al. [64]
	Extension	Choi et al. [125], Ferraro et al. [107]

important to consider the computational complexity of SHAP and its higher demand for computational resources.

E. AM-BASED POST HOC INTERPRETATION

AM is designed as a technique to augment the performance of models by enhancing their capability to discern internal correlations and extract global information, skills that are crucial in IFD [23], [145]. Besides, AM has been explored for its potential to improve long-range information extraction and alleviate the issue of catastrophic forgetting [146]. This section delves into the role of AM as a post hoc interpretation method in IIFD, highlighting its significance in boosting model interpretability and effectiveness in diagnosing faults by focusing on the most relevant features within data.

1) A BRIEF INTRODUCTION TO AM

The attention module serves as a pivotal computational unit for analyzing long-range dependencies, using a query (Q) alongside key (K)-value (V) pairs, as illustrated in Figure 12. Initially, matrices W^Q, W^K, and W^V, which are randomly initialized, are used to compute Q, K, and V. The process then involves calculating correlation coefficients between features by applying an attention function to Q and K. Equations (8) and (9) depict the most frequently utilized functions for this purpose. The attention output is derived as a weighted sum of the values, with weights obtained through a softmax function (10) applied to the correlation coefficients, adjusted by $\sqrt{d_k}$.

Attention mechanisms are broadly categorized into three types: soft attention, hard attention, and self-attention. Soft attention offers a probabilistic approach to weighing the importance of different parts of the input data, providing a differentiable solution that integrates smoothly with neural network architectures. This flexibility allows for the entire data context to be considered in a weighted manner, contributing to gradient-based learning. In contrast, hard attention selects specific segments of the input data to focus on, effectively ignoring the rest. This selection is non-differentiable, often relying on reinforcement learning techniques for optimization due to its discrete nature.

Self-attention stands apart by allowing input elements to directly interact and evaluate their mutual relevance, pinpointing focus areas within the data. It dynamically allocates attention based on the input's internal context, quantifying element interdependence. Self-attention's parallel computation capability is notably advantageous for lengthy sequences, performing simultaneous attention assessments across all input elements via straightforward matrix operations. Vaswani et al. [147] enhanced this with the introduction of multi-head attention, conducting attention operations in parallel across multiple "heads." This facilitates the model's ability to simultaneously attend to information from varied representational subspaces and positions. An example of this enhancement is demonstrated through an 8-head self-attention mechanism, detailed in Figure 13.

$$\text{dot-product} : \text{Similarity}(Query, Key_i) = Query \cdot Key_i. \tag{8}$$

$$\text{Cosin} : \text{Similarity}(Query, Key_i) = \frac{Query \times Key_i}{\|Query\| \times \|Key_i\|}. \tag{9}$$

$$\text{Softmax}(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}} \tag{10}$$

2) APPLICATION OF AM TO IIFD

AM have been increasingly applied in IIFD, as evidenced by several key studies [20], [148], [149], [150], [151]. These investigations primarily leverage AM for feature extraction and fault diagnosis but often do not address model interpretability. A comprehensive summary of AM applications in post-hoc IIFD is detailed in Table 9, categorizing AM-based IIFD into three distinct groups: CNN-based, Transformer-based, and Attention-based interpretability.

a: CNN INTEGRATION WITH AM

Combining AM with CNNs harnesses CNN's self-learning capacity and AM's proficiency in identifying key features. This combination facilitates attention score visualization, aiding the interpretation of diagnostic results. Li et al. [152] integrated AM with CNN to highlight important data segments and extract unique features for enhanced interpretation of diagnostic outcomes through attention visualization. Similarly, Wang et al. [153] examined CNN's feature-learning mechanism with AM's aid, offering insights into CNN model interpretability. Chan and Shuai [154] utilized AM to extract frequency-domain data features, enabling early degradation detection and component fault identification through attention-weight distribution.

b: TRANSFORMERS AND AM

Transformers, employing AM extensively and discarding conventional convolutional structures, achieve precise sequence-to-sequence (seq-to-seq) predictions through an encoder-decoder architecture built on AM layers. Transformers excel at capturing global associations with self-attention, yet they may struggle with clearly establishing the causal link between signal patterns and fault types. Addressing this challenge, Li et al. [155] introduced a variational attention-based transformer network for efficient association extraction in rotating machinery fault diagnosis. Tang et al. [156] proposed a signal transformer that explores signal state features across various spaces, enhancing model interpretability with an attention visualization approach for fault identification.

c: ATTENTION-BASED INTERPRETABILITY

Several innovative Attention-based IIFD methods aim to improve AM's efficacy. Sun et al. [157] employed an enhanced NonLocal-Pooling-Attention module for effective feature capture under noise, analyzing the model's internal workings through visualization. Liao et al. [158] derived an attention mechanism from quadratic neurons, offering inherent interpretability. Liu et al. [159] recommended an attention fusion unit for interpretable feature capture, visualizing attention weights to identify key time-domain signal components. Zhang et al. [110] introduced an attention-based network for feature extraction, addressing overfitting and highlighting essential information. Additionally, Zhang et al. [160] innovatively merged causal discovery with AM, enhancing model generalizability and interpretability by learning real causal connections between faults and symptoms.

It can be noted that the AM serves as an effective tool for enhancing the interpretability of models. Its capability to highlight significant features while ignoring irrelevant ones enables AM to provide explanations at the feature level. While some studies categorize AM as a form of ante-hoc interpretability due to its intuitive explanations, it is important to recognize that AM does not incorporate prior knowledge inherently. Instead, its attention scores are developed through iterative training processes, positioning AM more accurately within the realm of post-hoc interpretability. In addition, there are notable drawbacks. Firstly, AM can sometimes assign high attention weights to segments that are not related to faults, thereby misleadingly emphasizing them in the analysis. Second, the majority of studies focusing on AM are primarily concerned with post-hoc interpretability analysis, indicating a lack of integration of reasonable prior knowledge during the construction phase of the model. These drawbacks underscore the need to view AM as a post-hoc interpretability tool, designed to enhance model transparency and understanding after the model has been trained, rather than as an inherent part of the model's initial design and development.

TABLE 9. The application of AM to IIFD.

Object	Reference	Architectures (based)
Bearing	Li et al. [152], Wang et al. [161], Wang et al. [153], Chan et al. [154]	CNN
	Li et al. [155], Tang et al. [156], Wang et al. [162]	Transformer
	Yang et al. [26], Liao et al. [158], Liu et al. [163]	Attention
Gears	Li et al. [155], Tang et al. [156]	Transformer
	Sun et al. [157], Liu et al. [159]	Attention
Others	Li et al. [70], Li et al. [164], Chen et al. [165], Huang et al. [166]	CNN
	Zhang et al. [110], Zhang et al. [160]	Attention

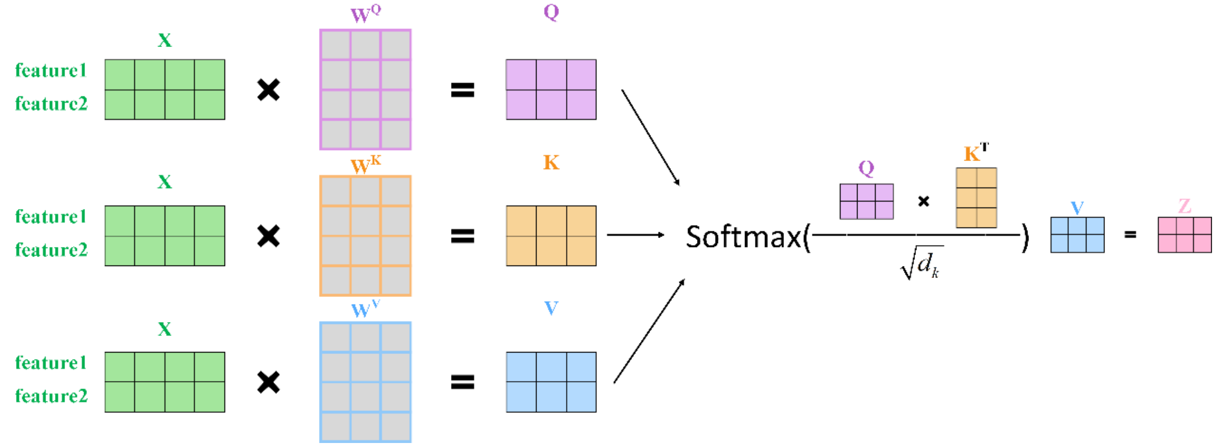


FIGURE 12. The calculation process of attention.

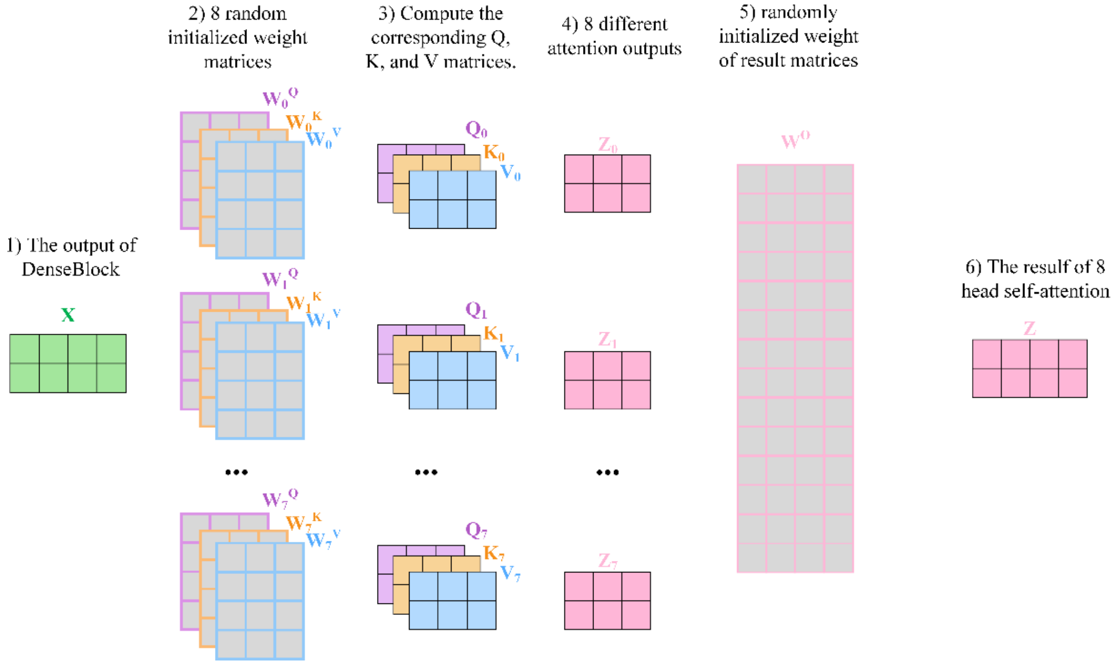

1) The output of DenseBlock
2) 8 random initialized weight matrices
3) Compute the corresponding Q, K, and V matrices.
4) 8 different attention outputs
5) randomly initialized weight of result matrices
6) The result of 8 head self-attention
X
feature1
feature2
W_0^Q
W_0^K
W_0^V
W_1^Q
W_1^K
W_1^V
W_7^Q
W_7^K
W_7^V
Q_0
K_0
V_0
Q_1
K_1
V_1
Q_7
K_7
V_7
Z_0
Z_1
Z_7
W^O
Z
...

FIGURE 13. The intuitive example of an 8 heads self-attention.

F. OTHERS APPROACHES

In addition to the methods outlined previously, alternative approaches exist employed in the post-hoc interpretable IIFD domain. We will present them in this subsection.

1) LAYER-WISE RELEVANCE PROPAGATION VISUALIZATION

Layer-Wise Relevance Propagation (LRP) is a commonly used visualization method to complete classification tasks [167]. As shown in Figure 14, The fundamental idea of LRP is to decompose the model prediction results and propagate the Relevance Score from the model output backward to the first layer.

Studies on LRP-based mechanical fault diagnosis mainly focus on bearings, gears, and motors. Grezmak et al. [168], [169], [170] studied the training performance of neural networks for motor vibration signals, gearbox fault types, and severity through the application of LRP and relevance score heatmaps. Kim et al. [171] used LRP in conjunction with the signal-preprocessing method for bearing IIFD in changing working conditions. Han et al. [172] confirmed that the current signal of a motor can be used in its deep fault condition by comparing the classical ideal feature point and the XAI-LRP output. Herwig et al. [173] analyzed the gear wear mechanism by applying LRP in tribological image training. Nie and Xie [174] proposed a normalized recurrent neural network for the early fault diagnosis of wind turbines and adopted LRP to reveal the model's decisions. Parziale et al. [175] investigated the diagnostic performance of LRP in condition monitoring of rotating shafts by examining correlation

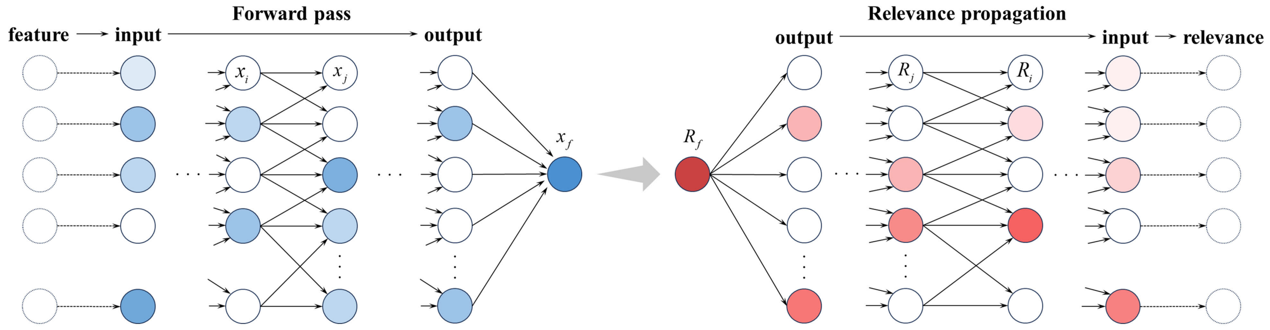


FIGURE 14. Illustration of the LRP: the left side represents a neural network's forward pass, processing features to produce an output. On the right, the output's relevance is traced backward through the network to the input features. This backward pass assigns relevance scores to each neuron, demonstrating their contribution to the final decision, making the model's internal reasoning transparent.

heatmaps. Based on LRP, Pan et al. [176] considered the effect of nonlinear activation functions, proposed Layer-wise contribution-filtered propagation. In addition, LRP is also used for Predictive maintenance decision interpreting [177], quantifying desirable properties in PHM [178], and explaining the relevance of inputs [179], [180], or trained weights in process control [181].

LRP-based post-hoc interpretable fault diagnosis models offer a powerful means to visually decode the decision-making process of a model. However, some disadvantages are preventing it from being applied by most researchers in mechanical fault diagnosis. Firstly, although the relevant heatmaps could visibly classify the inputs, they lack interpretability for the model itself, which is now of great importance in the fault diagnosis field [182]. Secondly, another crucial drawback of LRP is its strict criteria for input data, which poses a significant challenge in cases where IIFD problems involve data with a high signal-to-noise ratio.

2) SQUARE ENVELOP SPECTRUM

The Square Envelope Spectrum (SES) is an advanced method used for mechanical fault diagnosis. It operates by transforming the target signal into the Hilbert envelope spectrum through Hilbert and Fourier transforms. By analyzing the envelope spectrum, we can extract valuable information that may be challenging to detect in both the time and frequency domains.

In IIFD, SES has undergone significant enhancements, leading to improved interpretability in fault analysis. This enhanced SES technique not only facilitates fault diagnosis but also serves as a benchmark for evaluating the interpretability performance of other diagnostic methods. Based on the normalized square envelope spectrum [183], Hou et al. [184] designed an interpretable weights upgrade algorithm as an optimized square envelope spectrum to enhance the identification of the fault frequency. Wang et al. [9] introduced an approach to enhance the interpretability of the extreme learning machine by incorporating a square envelope and Fourier transform as input features. Ding et al. [185] employed the short-time Fourier transform and envelope spectrum analysis techniques to enhance the interpretability of the supposed deep learning-based method. Li et al. [8] corroborated the physical significance of the features extracted by their proposed model utilizing means of analyzing SES. Algburi et al. [186] utilized SES to evaluate the performance of the interpretable elements separated by singular spectrum analysis and generalized structured shrinkage algorithm. Yang et al. [187] applied SES in conjunction with the nonnegative matrix factorization method to accurately identify unknown fault modes in planetary gearbox systems. Li et al. [152] incorporated SES as an auxiliary tool to unravel the interpretability of attention mechanisms in the context of diagnostic applications.

It is noteworthy that the SES is frequently employed as a supplementary instrument for retrospective analysis in the realm of interpretable fault diagnosis. This auxiliary utilization of SES arises from its characteristics, facilitating its application in the examination and comprehension of diagnostic outcomes. However, its deployment primarily occurs subsequent to the initial diagnostic phase and mainly provides post-hoc interpretable results, and relying on prior knowledge leads to its limited application.

3) GRADIENT-BASED METHODS

Gradient-based methods are widely employed in the field of IIFD, including integrated gradients, gradient ascent, and gradient boosting. These methods analyze the gradients of the output relative to the input to quantify the impact of input features on model predictions, thereby elucidating feature importance and aiding in understanding model decision-making processes.

a: INTEGRATED GRADIENTS

Sipple [188] utilized IG to differentiate samples for anomaly interpretation in the Internet of Things. Peng et al. [189] introduced Smooth Integrated Gradients, which not only pinpoint responsible variables for faults but also provide a denoising effect in feature importance assessments.

Du et al. [190] applied IG in conjunction with continuous wavelet transform to select significant frequency ranges in frequency component analysis.

b: GRADIENT ASCENT (GA)

In the context of IIFD, Guo et al. [53] explored the workings of CNNs through frequency domain GA-based kernel visualization techniques, offering insights into the internal mechanisms of these networks.

c: GRADIENT BOOSTING AND XGBoost

XGBoost, an advanced form of gradient boosting that integrates decision tree boosting [191], is particularly effective in classifying features and determining their importance, embodying a post-hoc interpretability approach. It has been utilized for providing interpretable root causes in PHM [192] and for feature replacement in process control [193].

d: OTHER APPLICATIONS

An et al. [194] leveraged gradient mapping to enhance soft thresholding algorithms, establishing more interpretable classification criteria. Li et al. [195] developed a high-sensitivity gradient-based interpretation method that improves upon previous visualization techniques in terms of efficiency and accuracy.

Despite the sensitivity and accuracy of gradient-based methods in IIFD, they face challenges, particularly in parameter selection, such as setting the appropriate gradient threshold, which can significantly affect the outcomes. Additionally, these methods often assume linear changes in fault characteristics, which may lead to less than optimal performance in nonlinear scenarios, highlighting a critical area for further research and methodological refinement.

4) CUSTOMIZING RELEVANCE

Customizing relevance-based methods tailor algorithms to emphasize the importance of specific input features or data points relative to the output, providing a focused insight into how particular factors influence model decisions in a defined context. These methods harness customized relevance indicators to ascertain feature significance without relying on complex algorithmic frameworks, thereby facilitating post-hoc interpretation through visual representations.

For instance, Malhi et al. [196] employed the Contextual Importance and Utility (CIU) to decipher black box models, showcasing visual classification outcomes by utilizing CIU's analytical capabilities. CIU distinguishes between Contextual Importance (CI), which links directly to inputs, and Contextual Utility, which relates to outputs. Oliveira et al. [197] tackled the limitations of Autoencoders in anomaly detection by introducing the Residual eXPlainer, which computes feature correlation R_{nm} using Z-scores and provides explanations via deviation analysis of reconstructed input features. Zhuo et al. [198] drew inspiration from adversarial attacks to propose an adversarial fault reconstruction method, incorporating it into an explanatory framework that assesses variable contributions within general fault detection and classification models. Li et al. [28] defined three metrics to quantify the interpretability of deep neural networks, achieving a detailed activation map that offers enhanced resolution and better explainability of model results.

These methods are particularly effective when applied to specific operational conditions, as they are more targeted compared to general approaches in experimental scenarios. They also tend to have simpler structures, requiring fewer computational resources. However, designing a suitable relevance score algorithm that aligns with specific problems can be challenging, which may limit the generalizability of these methods. This often necessitates a balance between customized specificity and broad applicability in the design of relevance-based interpretative machine learning methods.

G. EPILOG

This section introduces post-hoc interpretable fault diagnosis methods, which rely on visual analysis of model classification results to provide explanations. However, these methodologies have two primary limitations in their applicability. Firstly, they are typically considered model-agnostic methods, which means they do not depend on specific knowledge of the model. However, in the context of IIFD problems, a sufficient understanding of the model is necessary. Secondly, the process of post-hoc fault diagnosis often neglects the incorporation of physical information related to the specific problem. This omission poses a risk of providing unfaithful explanations, rather than genuine knowledge derived from the data [199].

IV. ANTE-HOC INTERPRETATION OF IFD

While post-hoc interpretability methods provide valuable insights into model behavior, they inherently possess several limitations. First, although these methods attempt to project feature maps back to the input space for easier interpretation, the physical significance of these feature maps often remains obscure. Second, the training process involves convolutional kernels that are randomly initialized and optimized based on classification loss, which keeps the process largely opaque.

In this section, we examine scholarly works that address fault diagnosis issues through the lens of model intrinsic interpretability. The discussion is structured into three distinct subsections: 1) Model Embedding Interpretability, 2) Model Framework Interpretability, 3) Model Parameters Interpretability. These categories underscore different aspects of interpretability, each vital for a comprehensive understanding of how models diagnose faults and the transparency of their operations.

A. OVERVIEW

Interpretability, often synonymous with transparency, is an inherent characteristic of a model that denotes the clarity and understandability of its decisions and functions to humans [200]. To distinguish between post-hoc and intrinsic interpretability, Vollert et al. [201] categorized model transparency into three levels: simulatability, decomposability, and algorithmic transparency. A model is considered intrinsically interpretable if it satisfies the criteria at the decomposability level, where the features employed are inherently interpretable. Models built on non-interpretable features do not meet this criterion [202]. It is important to note that interpretable features are those embedded with prior knowledge or physical significance, differentiating them from features merely highlighted in post-hoc explanations. Furthermore, ante-hoc interpretability inherently encompasses aspects of post-hoc interpretability [203], [204].

Ante-hoc interpretability can be achieved through the use of straightforward, self-explanatory models or by integrating interpretability into a model's structure before training [205]. Simple models such as Logistic Regression, Decision Trees, and Rule-based systems are naturally interpretable and categorized under ante-hoc interpretability [201]. Integrating interpretability into a model's structure involves incorporating physical elements that enhance transparency. Terms frequently used to describe models with inherent interpretability include transparent models [200], intrinsic interpretability [206], ad-hoc interpretability [207], and ante-hoc interpretability [202].

Recent research in ante-hoc IIFD is expanding rapidly. We have collected over 30 articles across three domains—Fault Diagnosis, PHM, and Anomaly Detection—to analyze and review, summarized under the following categories:

1) Model Embedded Interpretability: Incorporation of physical technologies into model design, such as specific kernels, interpretable feature extraction model or expert defined signal processing approaches.

2) Model Framework Interpretability: Development of a diagnostic pipeline that integrates signal processing methods with data-driven models.

3) Model Parameters Interpretability: Implementation of sparse parameters or weight to imbue diagnostic models with physical relevance and interpretability.

In the sections that follow, we will delve into fault diagnosis research from these three interpretive perspectives, providing a comprehensive review of the current landscape.

B. INTERPRETABLE MODEL EMBEDDING

Model embedding-based intrinsic interpretability refers to the design of the interpretable network by embedding prior physical knowledge into the network structure, achieving feature-level explanation.

1) INTERPRETABLE CONVOLUTION KERNEL

The main idea of interpretable convolutional kernels is to rely on manual experience to embed physically meaningful kernels instead of conventional convolution kernels [208]. For example, in vibration fault diagnosis, the wavelet transform is a renowned signal processing technique. It is capable of transforming a raw signal from the time domain into the time-frequency domain using a wavelet basis function. Thus, Lan et al. [43] and Li et al. [209] explored a Wavelet Kernel Network (WKN) where the first convolutional layer is replaced by a continuous wavelet convolutional layer. This layer utilizes parameterized wavelet dictionaries for the wavelet transform of the input signal, with only scale and translation parameters learned from the input. This method emphasizes extracting impactful components from raw signals in the first layer of WKN, enhancing physical interpretability and robustness to varied data.

Nevertheless, WKN is limited to using only a single type of wavelet kernel for extracting fault-related features. This necessitates the pre-selection of an appropriate wavelet basis function tailored to various datasets and working conditions. Furthermore, the reliance on a single wavelet kernel restricts its capability to effectively capture informative fault features, especially in complex fault scenarios like compound faults [210], [211], [212]. To overcome this issue, Jiang et al. [213] constructed the multi-wavelet kernel convolution (MWKC) layer through selected four wavelet functions and utilized it to replace the first layer of the CNN to capture the different impulse excitations from raw vibration signals. To balance the varying significance of each wavelet kernel within the MWKC layer, a kernel weight recalibration module is devised to dynamically assign different weights to various wavelet convolutional kernels. The proposed Multi-Wavelet Kernel Convolution Neural Network incorporates the mechanical knowledge of fault impulse excitation into the CNN model to enhance interpretability and credibility. Recent studies employing wavelet convolution for ante-hoc interpretability are summarized in Table 10. In addition to wavelet-based kernels, other physics-based kernels are also proposed. Such as, Sadoughi and Hu [214] designed a physics-based kernel based on bearing fault characteristic frequencies and shaft speed to achieve the embedding of information pertinent to bearing-related fault features. Wu et al. [215] developed a multiple learnable multiplication filtering kernel which combined with a special antialiasing constraint join with L1 sparse regularization constraint to effectively separate fault features from complex spectrum in a comprehensible way.

Although interpretable convolution kernels have introduced some of the ante-hoc interpretability, the physical kernel model only replaced the first convolutional layer of the standard CNN and still cannot explain the whole network. Moreover, the analytical expression of the designed kernel has to be defined in advance by manual experience. In this way, the designed kernel is less adaptive so it can only extract a specific type of fault features. In addition, the replacement operation is only performed for some local layers, while the rest layers still follow conventional structures. Consequently, the interpretability of such models is still insufficient.

2) CUSTOM NETWORK STRUCTURES

To overcome the above problems, some custom network structures, designed to combine prior knowledge, are also being progressively developed for the implementation of physical knowledge embedding. A physics-based CNN was proposed in ref. [214], which added three signal processing techniques(spectral kurtosis, envelop analysis, and fast Fourier transform) to the front of the CNN as new layers to incorporate useful information from physical knowledge about bearings and their fault characteristics. Huang et al. [166] established a form of prior knowledge regarding the correlation between faults and attributes, utilizing the Pearson Correlation Coefficient. This knowledge was then integrated into the feature extractor of CNNs through AM. Consequently, the proposed attention-based CNN is designed to focus on the correlations between data regions and faults during the feature extraction process. Chen et al. [216] formulated the time-frequency network, where the physically meaningful time-frequency transform method is embedded into the traditional convolutional layer as a trainable preprocessing layer. This preprocessing layer named as time-frequency convolutional layer, is constrained by a well-designed kernel function to extract fault-related time-frequency information. It not only improves the diagnostic performance but also reveals the logical foundation of the CNN prediction in a frequency domain view.

Custom network structures to achieve ante-hoc interpretability of IIFD can help researchers understand the reasons behind model decisions to some extent, improving model reliability and the ability to handle issues with small sample sizes. When the model exhibits anomalies, physical knowledge can also assist in analyzing the problems. However, custom network structures increase the complexity of the model, making the design, training, and tuning more challenging. It also limits the flexibility of the model and poses a risk of overfitting, thereby reducing the model's generalization capability.

3) DOMAIN KNOWLEDGE EMBEDDED FRAMEWORK

In this subsection, we streamline various IIFD frameworks that embed domain expertise for fault diagnosis. The signal processing informed neural network (SPINN) framework, as introduced by Shang et al. [217], melds signal processing with deep learning through its Denoising Fault-Aware Wavelet Network. This network adopts a Wavelet transform for initial input processing, enhances feature extraction via thresholding in noisy conditions, and isolates pivotal fault features with index-based filtering before classification, prizing explainability and noise reduction. The power-perturbation-based decision boundary analysis by Gwak et al. [218] analyzes vibration classification models' decision boundaries through power variations in key frequency bands. It gauges frequency significance and model power sensitivity by testing with perturbed data, elucidating decision boundaries. Kim et al. [171] crafted the single domain generalizable and physically interpretable framework that synergizes signal preprocessing and neural networks. It integrates specific knowledge about impulse excitation signals to provide domain generalization and physical interpretation. Bayesian Networks are deployed for causal fault analysis, with Nor et al. [219] utilizing a Bayesian deep learning model with SHAP for transparent anomaly detection and prognostics. Yang et al. [220] employed a hybrid Bayesian Network that combines data and expert knowledge to delineate process variable interactions, thereby facilitating fault detection with graphically interpretable results. Lastly, addressing the imbalance in sample distribution, Liu et al. [163] introduced the Adversarial Variational Autoencoder with Sequential Attention (AVAE-SQA) for interpretable data augmentation in rolling bearing fault diagnosis. This approach incorporates variational inference and attention mechanisms, providing theoretical explanations of data distributions and decision rationales aligned with fault mechanisms.

Each framework embodies a strategic blend of domain knowledge and innovative machine learning techniques to enhance diagnostic precision and interpretability, pivotal for the application in real-world industrial settings.

C. INTERPRETABLE MODEL FRAMEWORK

To advance ante-hoc IIFD, interpretability model frameworks are developed as structured methodologies that enhance the clarity of machine learning models. Compared with previous mentioned IIFD methods, these frameworks facilitate an understanding of how inputs are transformed into outputs and help explain the decision-making processes involved. Given the diversity of theories behind each explainable framework, we provide detailed descriptions of select frameworks that demonstrate notable features.

1) ALGORITHM UNROLLING

Recently, algorithm unrolling [221] has garnered increasing interest in the DL, presenting an innovative solution to the challenge of model interpretability. This method involves unrolling each iteration of an iterative algorithm into a discrete network unit, based on the iterative formula. Then, these units are systematically connected to form a neural network. In this process, the predefined parameters of the iterative algorithm are reconfigured as adaptable parameters within the neural network. This allows for their optimization through backpropagation in an end-to-end manner, enhancing the network's learning capability. An algorithm unrolling network is interpretable since it is not empirically designed but is methodically developed under the procedure of an iterative algorithm [208]. Unrolled iterative algorithms, including the iterative shrinkage-thresholding algorithm (ISTA) [222], [223], [224], orthogonal matching pursuit [225], and the alternating direction method of multipliers [226]. In the domain of mechanical fault diagnosis, the exploration of algorithm unrolling networks is currently at an early stage.

Algorithm unrolling has rapidly advanced in theoretical research and practical applications in recent years. As a commonly used signal processing tool, sparse coding generally regards sparse coefficients as the features to characterize signals, which has been widely used in mechanical fault feature extraction [227]. The ISTA for the convolutional sparse coding optimization problem was unrolled into a neural network known as CISTA-Net by Rao et al. [208] using the algorithm unrolling approach. Because the iterative method determines the structure of CISTA-Net, it has a well-established theoretical foundation. Zhao et al. [228] introduced the layered general sparse coding (LGSC) algorithm, solving the general sparse coding issue with a multi-layered approach and evolving it into LGSC-Net using deep unrolling. This innovation bridges LGSC with CNNs by demonstrating how the soft nonnegative thresholding operator and the ReLU function, augmented with a bias term, are equivalently capable of representation. By integrating multi-layer processing, existing structures, and domain knowledge from the Multi-Layer Sparse Coding (ML-SC) model, LGSC demonstrates notable interpretability and a robust theoretical basis. Motivated by the ML-SC model, An et al. [194] developed a nested iterative soft thresholding algorithm (NISTA) as a solution for an ML-CSC model which is specifically designed for extracting fault features from vibration signals. To allow the parameters in the algorithm to adapt to various scenarios, NISTA is unrolled to form an explainable neural network. Within this network, all parameters are updated in an end-to-end manner using back-propagation. The design process of NISTA-Net is based on a well-defined theoretical framework, making the network's architecture both understandable and interpretable. Qin et al. [229] introduced a multi-scale component analysis network (MCAN), designed for high-performance and interpretable mechanical equipment fault diagnosis. MCAN is constructed by unrolling the iterative solution algorithm of a morphological component analysis (MCA) model. This model integrates multi-scale priors of vibration signals into a network, rendering the architecture network interpretable. Essentially, MCAN is an unrolled network of optimization algorithms tailored for the MCA model, incorporating vibration signal priors. The entire forward propagation process of the network is analogous to solving the MCA model's corresponding problem, thereby endowing the network with inherent interpretability.

Algorithm Unrolling transforms traditional iterative algorithms (such as optimization or signal processing algorithms) into an equivalent DL model. This approach combines the strengths of algorithms (like robustness and accuracy) with the adaptability and learning capabilities of deep learning. It enhances interpretability while improving efficiency and performance. Moreover, it serves as a bridge between traditional theoretical methods and data-based deep learning approaches. However, the design and implementation of algorithm unrolling is complex, requiring a deep understanding of both the original algorithm and deep learning architectures. Due to certain specific assumptions or limitations in traditional algorithms, it can be challenging to train with limited data. Similar to embedding physical knowledge, Algorithm Unrolling faces issues with poor generalization, making it difficult to adapt to different tasks or data types.

2) LOGICAL NETWORK FOR FORMAL LANGUAGES INTERPRETATION

Logical inference involves deriving logical expressions that describe system properties from data [230]. It is often implemented in a formal language in the ante-hoc IIFD. The fault diagnosis construction procedure can be formulated as a language generation process and the formal languages can be seen as interpretable classifiers, which provide interpretability for the fault diagnosis procedure. Recently, there has been applying temporal-logic-based formal language to diagnose faults and obtain good performance. Chen et al. [231] proposed a temporal logic neural network (TLNN), in which the network can be described and interpreted as a weighted signal temporal logic. TLNN not only keeps the nice properties of traditional neuron networks but also provides a logical interpretation of itself with formal language. The result of experience with real data sets shows the embedded formal language of the neuron network can provide explanations about the decision process, thus achieving interpretable fault diagnosis. Tian et al. [232] adopted the weighted signal temporal logic (wSTL) as a formal language and proposed a temporal logic network (TLN) for interpretable fault diagnosis of rolling element bearings. To further validate the interpretability of the model, timed failure propagation graphs are used to describe the logical relationship and propagation between fault events in the time domain. Experimental results demonstrate TLN's ability to extract impulse fault patterns from signals, accurately describe fault events through learned wSTL formulas, and enhance understanding of fault events for non-expert individuals through TFPGs. However, signal temporal logic (STL) is relatively weak in resisting noise, while real systems often operate in noisy environments. Since the fault signals of many systems are contaminated by noise and can only be detected in the frequency domain. Hence, Chen et al. [233] proposed a novel formal language for fault diagnosis, called signal spectral logic (SSL), which is inspired by the signal temporal logic and defined over signals' spectral kurtosis. The SSL is suitable to describe the spectral properties of time-series data and diagnose the fault for rotational machines, thus providing interpretation for the fault diagnosis results, and is robust to noisy environments. Another challenge of applying formal language to fault diagnosis is to find the optimal formula (sentence). Kong et al. [230] tried all combinations of basic formulas according to a predefined order and selected the best one. Nevertheless, this method suffered from a combinatorial explosion issue. To reduce the computational complexity, the author in [7] formulated the formula generation problem as a Markov decision process and solved it with a reinforcement learning algorithm. Furthermore, formal languages-based approaches like frequency temporal logic [234], and shapelet temporal logic [235] have also been utilized in IIFD.

Designing an interpretable fault diagnosis framework typically involves merging traditional fault diagnosis techniques with modern data-driven methods (such as machine learning or deep learning). The aim is to provide insights into the causes and nature of faults while maintaining transparency in the diagnostic process. However, designing and implementing an efficient and interpretable fault diagnosis framework can be highly complex and costly. Moreover, designing an effective interpretable fault diagnosis framework may require in-depth domain expertise and experience. Finally, overemphasizing interpretability might compromise the model's accuracy or efficiency, and the explanations provided could be based on the perspective of the framework, which might not be entirely accurate or comprehensive.

D. INTERPRETABLE MODEL PARAMETERS

In addition to the approaches discussed in earlier sections, our literature review has uncovered unique ante-hoc IIFD approaches, such as the introduction of sparsity and the explanation of model weights. These methods endow models with interpretability by incorporating specific parameters designed for ante-hoc application.

1) SPARSITY IN MODEL LEARNING

Given that fault signals from rotating machines often exhibit sparse, non-Gaussian, and non-stationary characteristics, many studies have adopted sparsity to enable models to learn interpretable fault signal representations. There are two main reasons for introducing sparsity [236]: 1) Sparse weight distributions inherently offer clearer explanations as they concentrate weight energy, simplifying the assessment of contributions and importance. 2) Fault vibration signals typically exhibit square envelope spectra indicating cyclic fault frequencies, which vary among different faults. Sparsity helps to better capture and differentiate these fault features. For instance, Pu et al. [237] developed a restricted sparse frequency-domain space (RSFDS) for rolling bearing fault features (RBFFs), incorporating a multichannel fusion mechanism that maps RBFFs to RSFDS, thus enhancing physical clarity and interpretability. Ma et al. [238] introduced a sparsity-constrained GAN model, imposing sparsity during the model's training phases to foster the learning of explainable signal representations.

2) EXPLANATION OF MODEL WEIGHTS

By pre-setting neural network weights based on physical knowledge, the interpretability of decision processes is enhanced, making outcomes more reliable. Yan et al. [205] constructed an interpretable weight matrix, which interacts with the time-frequency diagram to track the degradation process. This matrix, embedded within a neural network, determines the initial weights between the network's input and hidden layers, while the subsequent layers are optimized through intelligent algorithms. This structure ensures that the extracted features reflect fault-related characteristics, enabling the network to distinguish between normal and abnormal patterns effectively, thereby achieving ante-hoc interpretable fault diagnosis. Combining model weight explanations with sparsity further enhances interpretability, as demonstrated by Yan et al. [236] who developed a weight-oriented optimization model driven by discrimination and sparsity.

E. EPILOG

This section reviews ante-hoc interpretability in IIFD, dividing it into three directions: 1) model embedding interpretability, 2) model framework interpretability, and 3) model parameters interpretability. By summarizing the advantages and disadvantages of these directions, it reveals three main issues. First, there is the complexity and high cost of design. Embedding prior knowledge often requires a solid foundation in the principles of fault diagnosis and an in-depth understanding of deep learning. Second, there is the issue of poor generalization. The ante-hoc interpretable models designed are mostly based on specific physical principles. However, in reality, faults often involve multi-physics field coupling, so models may fail when working conditions change or in the presence of noise interference, making them unsuitable for different tasks or data types. Lastly, there is the issue of insufficient performance. The reasoning process added to interpretable models increases the computational burden, involving more parameters and complex structures, which can make training more difficult, especially in resource-limited situations. Therefore, researching a lightweight ante-hoc interpretable fault diagnosis model that maintains both versatility and robustness while preserving model performance is urgently needed.

V. DISCUSSION: FUTURE CHALLENGES IN IIFD

With the advancement of the IIFD, DL has gradually replaced the conventional fault diagnosis pattern of signal processing, feature extraction, and fault identification. This evolution towards deep learning enables the automatic extraction of features and identification of equipment failures, significantly reducing the reliance on prior diagnostic knowledge and enhancing both the efficiency and accuracy of fault identification. Nonetheless, the ‘‘black box’’ nature of deep learning models poses challenges to their reliability and generalizability, bringing the issue of interpretability, especially in rotating machinery diagnosis, to the forefront of urgent issues to be addressed. At the end of this review, we highlight the challenges faced by IIFD, aiming to position and stimulate interest in the future development trends of the IIFD field over the next decade, encouraging anticipation of and engagement with the potential directions this field may take.

A. DEFINING AND EVALUATING THE INTERPRETABILITY OF IIFD METHODS

In the realm of IIFD, various methods have been developed to interpret the outcomes of intelligent diagnostic processes. Despite these advancements, a standard definition of interpretability within intelligent diagnostics remains elusive.

TABLE 10. Papers of ante-hoc interpretable methods.

Paper	IFD Task	Dataset	Prior Knowledge	Interpretable method
Interpretable Model Embedding				
[43]	FD	Other	Laplace wavelet	Wavelet convolutional Kernels
[213]			Continuous wavelet	Multi-Wavelet Kernel
[214]			Rotational speed & Fault characteristics	Physics-based Conv
[216]		CWRU & Other	Time-Frequency transform	Time-frequency Conv
[215]		Other	Fault signal sparsity	Interpretable sparse kernels
[239]		CWRU	Lifting wavelets	Smart lifting wavelet kernels
[166]	PHM	Other	Category-attribute correlations	Attention-based feature extractor
[209]			Wavelet	Wavelet convolutional Kernels
[240]	FD		Discrete wavelet transform	Wavelet packet kernel
[241]			Fault characteristics	Adaptive fault attention mechanism
[242]			Fault frequency bandwidth	Sinc filters
[243]	PHM		Variational Bayesian inferences	Structured-effect neural network
[163]	FD	Other	The failure mechanism of rolling bearings	Adversarial variational autoencoder
[171]			Signal processing knowledge	The single domain generalizable and physically interpretable framework
[217]		XJTU-SY & Other	Signal processing knowledge	Signal Processing Informed Neural Network framework
[218]	FD	CWRU&PU	Power perturbation	Power-perturbation-based decision boundary analysis framework
[219]	PHM	Other	Uncertainty measure of the Bayesian	Bayesian deep learning model and SHAP
[220]			Interactions between process variables	Bayesian Networks
[27]			Generalized additive models	Bayesian Networks
[244]			Expert knowledge	Explainable deep convolutional autoencoder
Interpretable Model Framework				
[194]		XJTU & SQI	The nested iterative soft thresholding algorithm	Algorithm Unrolling
[208]		SEU	Convolutional sparse coding	
[228]		Other	General sparse coding	
[229]		SQI	The iterative solution algorithm of a morphological component analysis	
[230]	FD	Other	Signal temporal logic	Formal languages
[231]	FD	Other	Weighted signal temporal logic	
[232]		CWRU&MFPT &Other		
[233]		Other	Signal spectral logic	
[234]		Other	Frequency-temporal-logic	
[235]		CWRU	Shapelet temporal logic	
Interpretable Model Parameters				
[205]	PHM	XJTU	Fault signal time-frequency characteristics	Physically interpretable weight matrix
[236]	PHM	CWRU&Other	Weighted square envelope spectrum of degradation feature	Weight sparsity
[237]	FD	CWRU&Other	Rolling bearing fault feature	Quadratic complex domain equation
[238]	FD	CWRU	Vibration signal generation mechanism	Network weights analyzing

Current interpretation methods struggle to clearly articulate the complex internal learning mechanisms and decision processes of deep learning models, often relying on subjective factors that may lead to inconsistent, contradictory, or even incorrect interpretations. This subjectivity underscores the need for a unified standard to assess interpretability, which would minimize the impact of subjective interpretations and ensure more consistent and reliable diagnostic outcomes.

To address this, there is a pressing need to design metrics that can quantitatively evaluate interpretability. This would facilitate the establishment of a standardized benchmark for interpretability, transcending the current qualitative assessments that vary widely across different methods. By developing robust metrics, the field can measure and compare the reliability of interpretations, enhancing the credibility and utility of diagnostic models.

B. IMPROVING GENERALITY IN IIFD METHODS

The generality of models within IIFD poses substantial challenges, particularly when these models are tailored to specific types of physical knowledge. While this specialization enhances model interpretability for certain fault diagnoses, it severely restricts their applicability across different contexts. Models optimized for narrowly defined problems may not perform adequately or provide meaningful interpretations when confronted with unfamiliar data distributions. This limitation highlights the necessity of developing flexible models capable of maintaining high interpretability across various scenarios and fault types. Enhancing model adaptability would ensure that IIFD advancements can be effectively applied in diverse diagnostic environments, thus broadening their practical impact.

C. BALANCING PERFORMANCE AND INTERPRETABILITY IN IIFD MODELS

Achieving a balance between interpretability and performance poses a significant challenge in IIFD. Post-hoc interpretation methods can provide insights after model training by analyzing feature importance, but the variability in these interpretations can affect the robustness and reliability of the results. On the other hand, ante-hoc methods, which incorporate prior knowledge directly into model structures, offer more stable explanations but can sometimes compromise model performance. These methods constrain the model to focus on inherently interpretable features, which may limit its learning potential and necessitate greater computational resources. The ongoing challenge lies in developing models that transparently articulate their decision-making processes while performing effectively across diverse and dynamic environments. Achieving this balance is critical for advancing fault diagnosis technologies that are both trustworthy and highly functional.

D. INTEGRATING DOMAIN KNOWLEDGE WITH IIFD METHODS

Integrating domain knowledge into IIFD methods is imperative to ensure the reliability and relevance of diagnostic models. Domain knowledge provides essential insights into system mechanics and failure modes, which are crucial for tailoring feature engineering and model tuning processes. However, converting this often tacit knowledge into a format usable by automated systems poses significant challenges, particularly with complex machine learning architectures like deep learning. Successfully integrating domain knowledge enhances model reliability, aligns with regulatory transparency requirements, and ensures that diagnostics are grounded in substantive expert understanding. Overcoming these integration challenges is crucial for developing diagnostic systems that are both effective and accepted within industry sectors.

E. IDENTIFYING CAUSAL RELATIONSHIPS FROM FAULT INTERPRETATIONS

Identifying causal relationships rather than mere correlations in fault diagnostics is essential for accurate fault identification and effective intervention. This distinction is crucial in safety-critical applications such as aerospace and automotive, where incorrect diagnostics can lead to catastrophic failures. Addressing this challenge involves refining diagnostic models to discern causal relationships from vast data sets, ensuring that the faults identified are genuinely responsible for observed issues. Establishing these relationships not only enhances the reliability of diagnostic models but also supports targeted and efficient corrective actions, ultimately leading to safer and more dependable system operations.

VI. CONCLUSION

In this paper, we review the applications of interpretable DL models in IFD, which can roughly divided into post-hoc interpretation and ante-hoc interpretation methods. Post-hoc interpretation methods refer to explaining the diagnostic results of the model after the training, primarily by assessing feature importance to analyze the relationship between inputs and outputs. Common post-hoc methods include CAM, LIME, SHAP, and the AM. Utilizing post-hoc interpretation to analyze fault diagnosis results can enhance model transparency and boost practitioners' confidence in the diagnostic outcomes. Nevertheless, post-hoc interpretation is uncertain, as the interpretive results may fluctuate with the model's training and potentially yield inconsistent explanations. To further advance toward more interpretable IFD, recent research has explored the embedding of physical knowledge as a strategy for achieving ante-hoc interpretability. This approach ensures more stable and reliable explanations by incorporating prior expert knowledge into the model from the beginning. We innovative categorize ante-hoc IFD into three approaches based on their method of knowledge embedding: 1)interpretable model embedding, where designing convolutional kernels with physical significance, 2)interpretable model frameworks, where facilitating an understanding of how inputs are transformed into outputs by integrating or embedding expert knowledge within the learning framework, and 3)interpretable model parameters, where explainable model parameters make the learning process partly interpretable). Although ante-hoc IFD offers more robust explanations, it may compromise model performance and suffer from limited generalizability, potentially restricting its application in industry. To bridge the gap, the derivation of dynamic equations through deep learning inference holds promise for establishing a reliable and interpretable fault diagnosis model. Finally, we discuss the challenges of IIFD, hoping to provide readers with clear research directions. This review is expected to systematically present the development of IIFD and provide valuable guidelines for future research in this field.

REFERENCES

[1] Y. Xiao, H. Shao, M. Feng, T. Han, J. Wan, and B. Liu, ''Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer,'' *J. Manuf. Syst.*, vol. 70, pp. 186–201, Oct. 2023.

[2] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, ''Deep residual shrinkage networks for fault diagnosis,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[3] X. Jin, F. Cheng, Y. Peng, W. Qiao, and L. Qu, ''Drivetrain gearbox fault diagnosis: Vibration- and current-based approaches,'' *IEEE Ind. Appl. Mag.*, vol. 24, no. 6, pp. 56–66, Nov. 2018.

[4] Y. Wang, P. W. Tse, B. Tang, Y. Qin, L. Deng, and T. Huang, ''Kurtogram manifold learning and its application to rolling bearing weak signal detection,'' *Measurement*, vol. 127, pp. 533–545, Oct. 2018.

[5] F. Cong, J. Chen, G. Dong, and M. Pecht, ''Vibration model of rolling element bearings in a rotor-bearing system for fault diagnosis,'' *J. Sound Vibrat.*, vol. 332, no. 8, pp. 2081–2097, Apr. 2013.

[6] F. Jia, Y. Lei, H. Shan, and J. Lin, ''Early fault diagnosis of bearings using an improved spectral kurtosis by maximum correlated kurtosis deconvolution,'' *Sensors*, vol. 15, no. 11, pp. 29363–29377, Nov. 2015.

[7] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, ''Applications of machine learning to machine fault diagnosis: A review and roadmap,'' *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587.

[8] T. Li, C. Sun, S. Li, Z. Wang, X. Chen, and R. Yan, ''Explainable graph wavelet denoising network for intelligent fault diagnosis,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8535–8548, Jun. 2024.

[9] D. Wang, Y. Chen, C. Shen, J. Zhong, Z. Peng, and C. Li, ''Fully interpretable neural network for locating resonance frequency bands for machine condition monitoring,'' *Mech. Syst. Signal Process.*, vol. 168, Apr. 2022, Art. no. 108673.

[10] Z. Li, Z. Wu, Y. He, and C. Fulei, ''Hidden Markov model-based fault diagnostics method in speed-up and speed-down process for rotating machinery,'' *Mech. Syst. Signal Process.*, vol. 19, no. 2, pp. 329–339, Mar. 2005.

[11] F. Lu, Q. Tong, Z. Feng, Q. Wan, G. An, Y. Li, M. Wang, J. Cao, and T. Guo, ''Explainable 1DCNN with demodulated frequency features method for fault diagnosis of rolling bearing under time-varying speed conditions,'' *Meas. Sci. Technol.*, vol. 33, no. 9, Sep. 2022, Art. no. 095022.

[12] M. Xia, T. Li, L. Xu, L. Liu, and C. W. de Silva, ''Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks,'' *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2018.

[13] Y.-K. Gu, X.-Q. Zhou, D.-P. Yu, and Y.-J. Shen, ''Fault diagnosis method of rolling bearing using principal component analysis and support vector machine,'' *J. Mech. Sci. Technol.*, vol. 32, no. 11, pp. 5079–5088, Nov. 2018.

[14] A. Sharma, R. Jigyasu, L. Mathew, and S. Chatterji, ''Bearing fault diagnosis using weighted K-nearest neighbor,'' in *Proc. 2nd Int. Conf. Trends Electron. Informat. (ICOEI)*, May 2018, pp. 1132–1137.

[15] C.-C. Wang, Y. Kang, P.-C. Shen, Y.-P. Chang, and Y.-L. Chung, ''Applications of fault diagnosis in rotating machinery by using time series analysis with neural network,'' *Exp. Syst. Appl.*, vol. 37, no. 2, pp. 1696–1702, Mar. 2010.

[16] Z. Chen, W. Qin, G. He, J. Li, R. Huang, G. Jin, and W. Li, ''Explainable deep ensemble model for bearing fault diagnosis under variable conditions,'' *IEEE Sensors J.*, vol. 23, no. 15, pp. 17737–17750, Aug. 2023.

[17] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, ''Convolutional neural network based fault detection for rotating machinery,'' *J. Sound Vibrat.*, vol. 377, pp. 331–345, Sep. 2016.

[18] Y. Xue, D. Dou, and J. Yang, ''Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine,'' *Measurement*, vol. 156, May 2020, Art. no. 107571.

[19] Z. Xu, C. Li, and Y. Yang, ''Fault diagnosis of rolling bearings using an improved multi-scale convolutional neural network with feature attention mechanism,'' *ISA Trans.*, vol. 110, pp. 379–393, Apr. 2021.

[20] X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, and L. Zhang, ''Fault diagnosis for small samples based on attention mechanism,'' *Measurement*, vol. 187, Jan. 2022, Art. no. 110242.

[21] R. Liu, B. Yang, E. Zio, and X. Chen, ''Artificial intelligence for fault diagnosis of rotating machinery: A review,'' *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.

[22] S. Zhang, L. Su, J. Gu, K. Li, L. Zhou, and M. Pecht, ''Rotating machinery fault detection and diagnosis based on deep domain adaptation: A survey,'' *Chin. J. Aeronaut.*, vol. 36, no. 1, pp. 45–74, Jan. 2023.

[23] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, ''Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application,'' *Measurement*, vol. 199, Aug. 2022, Art. no. 111594.

[24] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang, ''A review of the application of deep learning in intelligent fault diagnosis of rotating machinery,'' *Measurement*, vol. 206, Jan. 2023, Art. no. 112346.

[25] F. Jia, Y. Lei, N. Lu, and S. Xing, ''Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization,'' *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, Sep. 2018.

[26] Z.-B. Yang, J.-P. Zhang, Z.-B. Zhao, Z. Zhai, and X.-F. Chen, ''Interpreting network knowledge with attention mechanism for bearing fault diagnosis,'' *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106829.

[27] J. Yang, Z. Yue, and Y. Yuan, ''Noise-aware sparse Gaussian processes and application to reliable industrial machinery health monitoring,'' *IEEE Trans. Ind. Informat.*, vol. 19, no. 4, pp. 5995–6005, Apr. 2023.

[28] S. Li, T. Li, C. Sun, R. Yan, and X. Chen, ''Multilayer grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis,'' *J. Manuf. Syst.*, vol. 69, pp. 20–30, Aug. 2023.

[29] B. Wang, Y. Lei, N. Li, and N. Li, ''A hybrid prognostics approach for estimating remaining useful life of rolling element bearings,'' *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[30] H. Qiu, J. Lee, J. Lin, and G. Yu, ''Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics,'' *J. Sound Vibrat.*, vol. 289, nos. 4–5, pp. 1066–1090, Feb. 2006.

[31] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, and X. Chen, ''Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study,'' *ISA Trans.*, vol. 107, pp. 224–255, Dec. 2020.

[32] S. Shao, S. McAleer, R. Yan, and P. Baldi, ''Highly accurate machine fault diagnosis using deep transfer learning,'' *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, ''Learning deep features for discriminative localization,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-CAM: Visual explanations from deep networks via gradient-based localization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[35] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, ''Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[36] M. B. Muhammad and M. Yeasin, ''Eigen-CAM: Class activation map using principal components,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[37] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, ''Fault diagnosis using eXplainable AI: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data,'' *Exp. Syst. Appl.*, vol. 232, Dec. 2023, Art. no. 120860.

[38] C.-J. Lin and J.-Y. Jhang, ''Bearing fault diagnosis using a grad-CAM-based convolutional neuro-fuzzy network,'' *Mathematics*, vol. 9, no. 13, p. 1502, Jun. 2021.

[39] Y. Yoo and S. Jeong, ''Vibration analysis process based on spectrogram using gradient class activation map with selection process of CNN model and feature layer,'' *Displays*, vol. 73, Jul. 2022, Art. no. 102233.

[40] H. Lu, A. M. Bray, C. Hu, A. T. Zimmerman, and H. Xu, "An interpretable deep learning method for bearing fault diagnosis," 2023, *arXiv:2308.10292*.

[41] H.-Y. Chen and C.-H. Lee, "Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis," *IEEE Access*, vol. 8, pp. 134246–134256, 2020.

[42] M. Saeki, J. Ogata, M. Murakawa, and T. Ogawa, "Visual explanation of neural network based rotation machinery anomaly detection system," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2019, pp. 1–4.

[43] H. Lan, W. Li, J. Chen, K. Feng, and R. Huang, "Wavelet convolutional neural network with multilabel classifier: A compound fault diagnosis framework and its interpretability analysis," in *Proc. Int. Conf. Sensing, Meas. Data Anal. Era Artif. Intell. (ICSMD)*, 2022, pp. 1–6.

[44] S. Yu, M. Wang, S. Pang, L. Song, and S. Qiao, "Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network," *Measurement*, vol. 196, Jun. 2022, Art. no. 111228.

[45] D. Yang, H. R. Karimi, and L. Gelman, "An explainable intelligence fault diagnosis framework for rotating machinery," *Neurocomputing*, vol. 541, Jul. 2023, Art. no. 126257.

[46] B. Chen, T. Liu, C. He, Z. Liu, and L. Zhang, "Fault diagnosis for limited annotation signals and strong noise based on interpretable attention mechanism," *IEEE Sensors J.*, vol. 22, no. 12, pp. 11865–11880, Jun. 2022.

[47] M. S. Kim, J. P. Yun, and P. Park, "An explainable neural network for fault diagnosis with a frequency activation map," *IEEE Access*, vol. 9, pp. 98962–98972, 2021.

[48] A. Hu, J. Sun, L. Xiang, and Y. Xu, "Rotating machinery fault diagnosis based on impact feature extraction deep neural network," *Meas. Sci. Technol.*, vol. 33, no. 11, Nov. 2022, Art. no. 114004.

[49] T. Han, C. Liu, L. Wu, S. Sarkar, and D. Jiang, "An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems," *Mech. Syst. Signal Process.*, vol. 117, pp. 170–187, Feb. 2019.

[50] F. Feng, C. Wu, J. Zhu, S. Wu, Q. Tian, and P. Jiang, "Research on multitask fault diagnosis and weight visualization of rotating machinery based on convolutional neural network," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 42, no. 11, p. 603, Nov. 2020.

[51] C. He, H. Shi, J. Si, and J. Li, "Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings," *J. Manuf. Syst.*, vol. 70, pp. 579–592, Oct. 2023.

[52] K. Wen, R. Huang, D. Li, Z. Chen, and W. Li, "Gradient-based interpretable graph convolutional network for bearing fault diagnosis," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2023, pp. 1–6.

[53] L. Guo, X. Gu, Y. Yu, A. Duan, and H. Gao, "An analysis method for interpretability of convolutional neural network in bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.

[54] H. Sun, X. Cao, C. Wang, and S. Gao, "An interpretable anti-noise network for rolling bearing fault diagnosis based on FSWT," *Measurement*, vol. 190, Feb. 2022, Art. no. 110698.

[55] H. Hao, F. Fuzhou, Z. Junzhen, Z. Xun, J. Pengcheng, J. Feng, X. Jun, L. Yazhi, and S. Guanghui, "Research on fault diagnosis method based on improved CNN," *Shock Vibrat.*, vol. 2022, pp. 1–15, Dec. 2022.

[56] C. Liu, Y. Meerten, K. Declercq, and K. Gryllias, "Vibration-based gear continuous generating grinding fault classification and interpretation with deep convolutional neural network," *J. Manuf. Processes*, vol. 79, pp. 688–704, Jul. 2022.

[57] C. Oh and J. Jeong, "VODCA: Verification of diagnosis using CAM-based approach for explainable process monitoring," *Sensors*, vol. 20, no. 23, p. 6858, Nov. 2020.

[58] O. Mey and D. Neufeld, "Explainable AI algorithms for vibration data-based fault detection: Use case-adadpted methods and critical evaluation," *Sensors*, vol. 22, no. 23, p. 9037, Nov. 2022.

[59] H. Huh, S. Y. Lee, S. Lee, K. H. Sun, and J. H. Jung, "New way of detecting vibration of mechanical systems by explainable deep learning," in *Proc. INTER-NOISE NOISE-CON Congr. Conf.*, vol. 261, no. 1. Seoul, South Korea: Institute of Noise Control Engineering, 2020, pp. 5646–5650.

[60] J. Liu, L. Hou, R. Zhang, X. Sun, Q. Yu, K. Yang, and X. Zhang, "Explainable fault diagnosis of oil-gas treatment station based on transfer learning," *Energy*, vol. 262, Jan. 2023, Art. no. 125258.

[61] J. Liu, L. Hou, X. Wang, R. Zhang, X. Sun, L. Xu, and Q. Yu, "Explainable fault diagnosis of gas-liquid separator based on fully convolutional neural network," *Comput. Chem. Eng.*, vol. 155, Dec. 2021, Art. no. 107535.

[62] M. S. Kim, J. P. Yun, and P. Park, "An explainable convolutional neural network for fault diagnosis in linear motion guide," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4036–4045, Jun. 2021.

[63] T. Ren, T. Han, Q. Guo, and G. Li, "Analysis of interpretability and generalizability for power converter fault diagnosis based on temporal convolutional networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[64] Z. Li, Y. Zhang, J. Ai, Y. Zhao, Y. Yu, and Y. Dong, "A lightweight and explainable data-driven scheme for fault detection of aerospace sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 6, pp. 8392–8410, Dec. 2023.

[65] Q. Chao, X. Wei, J. Tao, C. Liu, and Y. Wang, "Cavitation recognition of axial piston pumps in noisy environment based on grad-CAM visualization technique," *CAAI Trans. Intell. Technol.*, vol. 8, no. 1, pp. 206–218, Mar. 2023.

[66] W. Cheng, S. Wang, Y. Liu, X. Chen, Z. Nie, J. Xing, R. Zhang, and Q. Huang, "A novel planetary gearbox fault diagnosis method for nuclear circulating water pump with class imbalance and data distribution shift," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[67] P. Dopierala, "Fault detection method for energy measurement systems equipped with a Rogowski coil using the coil's response to a unit voltage jump and a fully convolutional neural network," *Measurement*, vol. 190, Feb. 2022, Art. no. 110749.

[68] A. Nasiri, A. Taheri-Garavand, M. Omid, and G. M. Carlomagno, "Intelligent fault diagnosis of cooling radiator based on deep learning analysis of infrared thermal images," *Appl. Thermal Eng.*, vol. 163, Dec. 2019, Art. no. 114410.

[69] S. Lee, H. Yu, H. Yang, I. Song, J. Choi, J. Yang, G. Lim, K.-S. Kim, B. Choi, and J. Kwon, "A study on deep learning application of vibration data and visualization of defects for predictive maintenance of gravity acceleration equipment," *Appl. Sci.*, vol. 11, no. 4, p. 1564, Feb. 2021.

[70] M. Li, P. Peng, J. Zhang, H. Wang, and W. Shen, "SCCAM: Supervised contrastive convolutional attention mechanism for ante-hoc interpretable fault diagnosis with limited fault samples," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6194–6205, May 2024.

[71] C. Ardito, Y. Deldjoo, T. D. Noia, E. D. Sciascio, and F. Nazary, "Visual inspection of fault type and zone prediction in electrical grids using interpretable spectrogram-based CNN modeling," *Exp. Syst. Appl.*, vol. 210, Dec. 2022, Art. no. 118368.

[72] J. Liu, L. Hou, S. He, X. Zhang, Q. Yu, K. Yang, and Y. Li, "Two-dimensional explainability method for fault diagnosis of fluid machine," *Process Saf. Environ. Protection*, vol. 178, pp. 1148–1160, Oct. 2023.

[73] G. Li, Q. Yao, C. Fan, C. Zhou, G. Wu, Z. Zhou, and X. Fang, "An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems," *Building Environ.*, vol. 203, Oct. 2021, Art. no. 108057.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[75] W. Wang, P. Wei, H. Liu, C. Zhu, G. Deng, and H. Liu, "A micromechanics-based machine learning model for evaluating the microstructure-dependent rolling contact fatigue performance of a martensitic steel," *Int. J. Mech. Sci.*, vol. 237, Jan. 2023, Art. no. 107784.

[76] J. A. Recio-García, B. Díaz-Agudo, and V. Pino-Castilla, "CBR-LIME: A case-based reasoning approach to provide specific local interpretable model-agnostic explanations," in *Proc. 28th Int. Conf. Case-Based Reasoning*, Salamanca, Spain. New York, NY, USA: Springer, Jun. 2020, pp. 179–194.

[77] A. Saini and R. Prasad, "Select wisely and explain: Active learning and probabilistic local post-hoc explainability," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 599–608.

[78] M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," 2019, *arXiv:1906.10263*.

[79] Z. Dikopoulou, S. Moustakidis, and P. Karlsson, "GLIME: A new graphical methodology for interpretable model-agnostic explanations," 2021, *arXiv:2107.09927*.

[80] X. Xiang, H. Yu, Y. Wang, and G. Wang, "Stable local interpretable model-agnostic explanations based on a variational autoencoder," *Int. J. Speech Technol.*, vol. 53, no. 23, pp. 28226–28240, Dec. 2023.

[81] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*.

[82] A. L. Alfeo, M. G. C. A. Cimino, and G. Vaglini, "Degradation stage classification via interpretable feature learning," *J. Manuf. Syst.*, vol. 62, pp. 972–983, Jan. 2022.

[83] D. C. Sanakkayala, V. Varadarajan, N. Kumar, G. Soni, P. Kamat, S. Kumar, S. Patil, and K. Kotecha, "Explainable AI for bearing fault prognosis using deep learning techniques," *Micromachines*, vol. 13, no. 9, p. 1471, Sep. 2022.

[84] M. T. Pham, J.-M. Kim, and C. H. Kim, "Rolling bearing fault diagnosis based on improved GAN and 2-D representation of acoustic emission signals," *IEEE Access*, vol. 10, pp. 78056–78069, 2022.

[85] A. Amin, A. Bibo, M. Panyam, and P. Tallapragada, "Vibration based fault diagnostics in a wind turbine planetary gearbox using machine learning," *Wind Eng.*, vol. 47, no. 1, pp. 175–189, Feb. 2023.

[86] A. Amin, A. Bibo, M. Panyam, and P. Tallapragada, "Wind turbine gearbox fault diagnosis using cyclostationary analysis and interpretable CNN," *J. Vibrat. Eng. Technol.*, vol. 12, no. 2, pp. 1695–1705, Feb. 2024.

[87] K. T. Mai, T. Davies, L. D. Griffin, and E. Benetos, "Explaining the decision of anomalous sound detectors," in *Proc. 7th Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Nancy, France, Nov. 2022, pp. 1–5.

[88] C. Yao, X. Yueyun, C. Jinwei, and Z. Huisheng, "A novel gas path fault diagnostic model for gas turbine based on explainable convolutional neural network with lime method," in *Proc. Turbo Expo, Power Land, Sea, Air*, vol. 84966, 2021, Art. no. V004T05A008.

[89] S. Srinivasan, P. Arjunan, B. Jin, A. L. Sangiovanni-Vincentelli, Z. Sultan, and K. Poolla, "Explainable AI for chiller fault-detection systems: Gaining human trust," *Computer*, vol. 54, no. 10, pp. 60–68, Oct. 2021.

[90] U. E. Akpudo and J.-W. Hur, "An explainable DL-based condition monitoring framework for water-emulsified diesel CR systems," *Electronics*, vol. 10, no. 20, p. 2522, Oct. 2021.

[91] W. Yu, X. Li, and Y. Sun. (Jul. 2021). *Towards Making Predictive Maintenance System Adaptive and Interpretable*. [Online]. Available: https://www.preprints.org/manuscript/202107.0040/v1

[92] D. Kim and J. Lee, "Predictive evaluation of spectrogram-based vehicle sound quality via data augmentation and explainable artificial intelligence: Image color adjustment with brightness and contrast," *Mech. Syst. Signal Process.*, vol. 179, Nov. 2022, Art. no. 109363.

[93] M. Yang, C. Xu, Y. Bai, M. Ma, and X. Su, "Investigating black-box model for wind power forecasting using local interpretable model-agnostic explanations algorithm: Why should a model be trusted?" *CSEE J. Power Energy Syst.*, early access, Jan. 25, 2023.

[94] G. Protopapadakis, A. Apostolidis, and A. I. Kalfas, "Explainable and interpretable AI-assisted remaining useful life estimation for aeroengines," in *Proc. Turbo Expo, Power Land, Sea, Air*, vol. 85987, 2022, Art. no. V002T05A002.

[95] A. Udo Sass, E. Esatbeyoglu, and T. Iwwerks, "Signal pre-selection for monitoring and prediction of vehicle powertrain component aging," *Sci. Technique*, vol. 18, no. 6, pp. 519–524, Dec. 2019.

[96] K. Kobayashi and S. B. Alam, "Explainable, interpretable, and trustworthy AI for an intelligent digital twin: A case study on remaining useful life," *Eng. Appl. Artif. Intell.*, vol. 129, Mar. 2024, Art. no. 107620.

[97] O. Serradilla, E. Zugasti, C. Cernuda, A. Aranburu, J. R. de Okariz, and U. Zurutuza, "Interpreting remaining useful life estimations combining explainable artificial intelligence and domain knowledge in industrial machinery," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–8.

[98] M. Baptista, M. Mishra, E. Henriques, and H. Prendinger. (2020). *Using Explainable Artificial Intelligence to Interpret RemainingUseful Life Estimation With Gated Recurrent Unit*. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-80191

[99] M. Al-Zeyadi, J. Andreu-Perez, H. Hagras, C. Royce, D. Smith, P. Rzonsowski, and A. Malik, "Deep learning towards intelligent vehicle fault diagnosis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–7.

[100] T. Khan, K. Ahmad, J. Khan, I. Khan, and N. Ahmad, "An explainable regression framework for predicting remaining useful life of machines," in *Proc. 27th Int. Conf. Autom. Comput. (ICAC)*, 2022, pp. 1–6.

[101] M. H. Widianto, A. A. S. Gunawan, Y. Heryadi, and W. Budiharto, "Evaluation of machine learning on smart home data for prediction of electrical energy consumption," in *Proc. Int. Conf. Comput. Sci., Inf. Technol. Eng. (ICCoSITE)*, Feb. 2023, pp. 434–439.

[102] J. P. Usuga-Cadavid, S. Lamouri, B. Grabot, and A. Fortin, "Using deep learning to value free-form text data for predictive maintenance," *Int. J. Prod. Res.*, vol. 60, no. 14, pp. 4548–4575, Jul. 2022.

[103] M. Madhikermi, A. K. Malhi, and K. Främling, "Explainable artificial intelligence based heat recycler fault detection in air handling unit," in *Proc. Int. Workshop Explainable, Transparent Auto. Agents Multi-Agent Syst.*, Montreal, QC, Canada. New York, NY, USA: Springer, 2019, pp. 110–125.

[104] J. Sharma, M. L. Mittal, and G. Soni. (Apr. 2023). *Explainable Artificial Intelligence (XAI) Enabled Anomaly Detection and Fault Classification of an Industrial Asset*. [Online]. Available: https://www.researchsquare.com/article/rs-2780708/v1

[105] D. Liang and F. Xue, "Integrating automated machine learning and interpretability analysis in architecture, engineering and construction industry: A case of identifying failure modes of reinforced concrete shear walls," *Comput. Ind.*, vol. 147, May 2023, Art. no. 103883.

[106] S. Sairam, S. Seshadhri, G. Marafioti, S. Srinivasan, G. Mathisen, and K. Bekiroglu, "Edge-based explainable fault detection systems for photovoltaic panels on edge nodes," *Renew. Energy*, vol. 185, pp. 1425–1440, Feb. 2022.

[107] A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 7279–7314, Jul. 2023.

[108] P. Pandey, A. Rai, and M. Mitra, "Explainable 1-D convolutional neural network for damage detection using Lamb wave," *Mech. Syst. Signal Process.*, vol. 164, Feb. 2022, Art. no. 108220.

[109] R. Tang, S. Zhang, W. Wu, S. Zhang, and Z. Han, "Explainable deep learning based ultrasonic guided wave pipe crack identification method," *Measurement*, vol. 206, Jan. 2023, Art. no. 112277.

[110] H. Zhang, J. Lin, J. Hua, T. Zhang, and T. Tong, "Attention-based interpretable prototypical network towards small-sample damage identification using ultrasonic guided waves," *Mech. Syst. Signal Process.*, vol. 188, Apr. 2023, Art. no. 109990.

[111] A. Hanchate, S. T. S. Bukkapatnam, K. H. Lee, A. Srivastava, and S. Kumara, "Explainable AI (XAI)-driven vibration sensing scheme for surface quality monitoring in a smart surface grinding process," *J. Manuf. Processes*, vol. 99, pp. 184–194, Aug. 2023.

[112] D. M. Onchis and G.-R. Gillich, "Stable and explainable deep learning damage prediction for prismatic cantilever steel beam," *Comput. Ind.*, vol. 125, Feb. 2021, Art. no. 103359.

[113] S. Sairam, S. Srinivasan, G. Marafioti, B. Subathra, G. Mathisen, and K. Bekiroglu, "Explainable incipient fault detection systems for photovoltaic panels," 2020, *arXiv:2011.09843*.

[114] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[115] S. Asutkar and S. Tallur, "An explainable unsupervised learning framework for scalable machine fault detection in industry 4.0," *Meas. Sci. Technol.*, vol. 34, no. 10, Oct. 2023, Art. no. 105123.

[116] T. Son Pham, D. Huy Nguyen, and N. Duy Minh Phan, "An interpretable machine learning approach for fault classification in bearing systems," in *Proc. 7th Nat. Scientific Conf. Applying New Technol. Green Buildings (ATiGB)*, Nov. 2022, pp. 166–170.

[117] A. Sinha, S. F. Ahmed, and D. Das, "Explainable AI for bearing fault detection systems: Gaining human trust," in *Proc. IEEE Guwahati Subsection Conf. (GCON)*, Jun. 2023, pp. 1–6.

[118] E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," *Appl. Sci.*, vol. 13, no. 4, p. 2038, Feb. 2023.

[119] G. Daiki, I. Tsuyoshi, H. Takekiyo, Y. Shota, K. Keiichi, T. Shigeyuki, and H. Akira, "Failure diagnosis and physical interpretation of journal bearing for slurry liquid using long-term real vibration data," *Struct. Health Monitor.*, vol. 23, no. 2, pp. 1201–1216, Mar. 2024.

[120] P. Kumar and A. S. Hati, "Deep convolutional neural network based on adaptive gradient optimizer for fault detection in SCIM," *ISA Trans.*, vol. 111, pp. 350–359, May 2021.

[121] M. J. Hasan, M. Sohaib, and J.-M. Kim, "An explainable AI-based fault diagnosis model for bearings," *Sensors*, vol. 21, no. 12, p. 4070, Jun. 2021.

[122] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mech. Syst. Signal Process.*, vol. 163, Jan. 2022, Art. no. 108105.

[123] D. Yao, G. Li, H. Liu, and J. Yang, "An intelligent method of roller bearing fault diagnosis and fault characteristic frequency visualization based on improved MobileNet V3," *Meas. Sci. Technol.*, vol. 32, no. 12, Dec. 2021, Art. no. 124009.

[124] Y. Wang and P. Wang, "Explainable machine learning for motor fault diagnosis," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2023, pp. 1–6.

[125] S. H. Choi and J. M. Lee, "Explainable fault diagnosis model using stacked autoencoder and kernel shap," in *Proc. IEEE Int. Symp. Adv. Control Ind. Processes (AdCONIP)*, Aug. 2022, pp. 182–187.

[126] B. Steurtewagen and D. Van den Poel, "Adding interpretability to predictive maintenance by machine learning on sensor data," *Comput. Chem. Eng.*, vol. 152, Sep. 2021, Art. no. 107381.

[127] W. Kirchgässner, O. Wallscheid, and J. Böcker, "Estimating electric motor temperatures with deep residual machine learning," *IEEE Trans. Power Electron.*, vol. 36, no. 7, pp. 7480–7488, Jul. 2021.

[128] X. Gu, K. See, Y. Wang, L. Zhao, and W. Pu, "The sliding window and SHAP theory—An improved system with a long short-term memory network model for state of charge prediction in electric vehicle application," *Energies*, vol. 14, no. 12, p. 3692, Jun. 2021.

[129] S. B. Ramezani, A. Amirlatifi, T. Kirby, M. Seale, and S. Rahimi, "Explainable machinery faults prediction using ensemble tree classifiers: Bagging or boosting?" in *Proc. Annu. Conf. PHM Soc.*, 2021, vol. 13, no. 1, pp. 1–12.

[130] C. Hwang and T. Lee, "E-SFD: Explainable sensor fault detection in the ICS anomaly detection system," *IEEE Access*, vol. 9, pp. 140470–140486, 2021.

[131] Y. Fang, H. Min, X. Wu, X. Lei, S. Chen, R. Teixeira, and X. Zhao, "Toward interpretability in fault diagnosis for autonomous vehicles: Interpretation of sensor data anomalies," *IEEE Sensors J.*, vol. 23, no. 5, pp. 5014–5027, Mar. 2023.

[132] H. T. T. Nguyen, H. Q. Cao, K. V. T. Nguyen, and N. D. K. Pham, "Evaluation of explainable artificial intelligence: Shap, lime, and cam," in *Proc. FPT AI Conf.*, 2021, pp. 1–6.

[133] B. An, Y. Ha, Y. Lee, W. Kwak, and Y. Lee, "Investigation on feature attribution for remaining useful life prediction model of cryogenic ball bearing," in *Proc. Int. Conf. Rotor Dyn.* New York, NY, USA: Springer, 2023, pp. 291–299.

[134] O. T. Bindingsbø, M. Singh, K. Øvsthus, and A. Keprate, "Fault detection of a wind turbine generator bearing using interpretable machine learning," *Frontiers Energy Res.*, vol. 11, Dec. 2023, Art. no. 1284676.

[135] J. Hu, Y. Zhang, W. Li, X. Zheng, and Z. Tian, "Trustworthy artificial intelligence based on an explicable temporal feature network for industrial fault diagnosis," *Cognit. Comput.*, vol. 16, no. 2, pp. 534–545, Mar. 2024.

[136] N. Herwig and P. Borghesani, "Explaining deep neural networks processing raw diagnostic signals," *Mech. Syst. Signal Process.*, vol. 200, Oct. 2023, Art. no. 110584.

[137] H. Zhang, G. Grünert, M. Solf, J. Brimmers, S. Barth, and T. Bergs, "Cross-process chain analysis on gear quality and sustainability," in *Proc. Congr. German Academic Assoc. Prod. Technol.* New York, NY, USA: Springer, 2023, pp. 174–184.

[138] A. Movsessian, D. G. Cava, and D. Tcherniak, "Interpretable machine learning in damage detection using Shapley additive explanations," *ASCE-ASME J. Risk Uncertainty Eng. Syst., B, Mech. Eng.*, vol. 8, no. 2, Jun. 2022, Art. no. 021101.

[139] H. Sasaki and K. Yamamura, "Topology optimization with Shapley additive explanations for permanent magnet synchronous motors," *IEEE Trans. Magn.*, vol. 60, no. 3, pp. 1–4, Mar. 2024.

[140] A. L. O. Vitor, A. Goedtel, S. Barbon, G. H. Bazan, M. F. Castoldi, and W. A. Souza, "Induction motor short circuit diagnosis and interpretation under voltage unbalance and load variation conditions," *Exp. Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 119998.

[141] Z. Pan and S. Fang, "Torque performance improvement of permanent magnet arc motor based on two-step strategy," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7523–7534, Nov. 2021.

[142] G. Youness and A. Aalah, "An explainable artificial intelligence approach for remaining useful life prediction," *Aerospace*, vol. 10, no. 5, p. 474, May 2023.

[143] M. L. Baptista, K. Goebel, and E. M. P. Henriques, "Relation between prognostics predictor evaluation metrics and local interpretability SHAP values," *Artif. Intell.*, vol. 306, May 2022, Art. no. 103667.

[144] M. Szelazek, S. Bobek, A. Gonzalez-Pardo, and G. J. Nalepa, "Towards the modeling of the hot rolling industrial process. preliminary results," in *Proc. 21st Int. Conf.*, Guimaraes, Portugal. New York, NY, USA: Springer, Nov. 2020, pp. 385–396.

[145] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[146] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[147] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[148] Z. Long, X. Zhang, L. Zhang, G. Qin, S. Huang, D. Song, H. Shao, and G. Wu, "Motor fault diagnosis using attention mechanism and improved AdaBoost driven by multi-sensor information," *Measurement*, vol. 170, Jan. 2021, Art. no. 108718.

[149] S. Yang, X. Kong, Q. Wang, Z. Li, H. Cheng, and K. Xu, "Deep multiple auto-encoder with attention mechanism network: A dynamic domain adaptation method for rotary machine fault diagnosis under different working conditions," *Knowl.-Based Syst.*, vol. 249, Aug. 2022, Art. no. 108639.

[150] L. Xiang, P. Wang, X. Yang, A. Hu, and H. Su, "Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism," *Measurement*, vol. 175, Apr. 2021, Art. no. 109094.

[151] X. Kong, X. Li, Q. Zhou, Z. Hu, and C. Shi, "Attention recurrent autoencoder hybrid model for early fault diagnosis of rotating machinery," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[152] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.

[153] H. Wang, Z. Liu, D. Peng, and M. J. Zuo, "Interpretable convolutional neural network with multilayer wavelet for noise-robust machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 195, Jul. 2023, Art. no. 110314.

[154] Y.-L. Chan and H.-H. Shuai, "Explainable health state prediction for social IoTs through multi-channel attention," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[155] Y. Li, Z. Zhou, C. Sun, X. Chen, and R. Yan, "Variational attention-based interpretable transformer network for rotary machine fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6180–6193, May 2024.

[156] J. Tang, G. Zheng, C. Wei, W. Huang, and X. Ding, "Signal-transformer: A robust and interpretable method for rotating machinery intelligent fault diagnosis under variable operating conditions," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[157] H. Sun, C. Wang, and X. Cao, "An adaptive anti-noise gear fault diagnosis method based on attention residual prototypical network under limited samples," *Appl. Soft Comput.*, vol. 125, Aug. 2022, Art. no. 109120.

[158] J.-X. Liao, H.-C. Dong, Z.-Q. Sun, J. Sun, S. Zhang, and F.-L. Fan, "Attention-embedded quadratic network (Qttention) for effective and interpretable bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[159] S. Liu, J. Huang, J. Ma, and J. Luo, "SRMANet: Toward an interpretable neural network with multi-attention mechanism for gearbox fault diagnosis," *Appl. Sci.*, vol. 12, no. 16, p. 8388, Aug. 2022.

[160] C. Zhang, X. Tian, Y. Zhao, T. Li, Y. Zhou, and X. Zhang, "Causal discovery-based external attention in neural networks for accurate and reliable fault detection and diagnosis of building energy systems," *Building Environ.*, vol. 222, Aug. 2022, Art. no. 109357.

[161] H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5735–5745, Sep. 2020.

[162] G. Wang, D. Liu, and L. Cui, "Auto-embedding transformer for interpretable few-shot fault diagnosis of rolling bearings," *IEEE Trans. Rel.*, vol. 73, no. 2, pp. 1270–1279, Jun. 2024.

[163] Y. Liu, H. Jiang, R. Yao, and H. Zhu, "Interpretable data-augmented adversarial variational autoencoder with sequential attention for imbalanced fault diagnosis," *J. Manuf. Syst.*, vol. 71, pp. 342–359, Dec. 2023.

[164] S. Li, J. Luo, and Y. Hu, "Toward interpretable process monitoring: Slow feature analysis-aided autoencoder for spatiotemporal process feature learning," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[165] S. Chen, L. Luo, Q. Xia, and L. Wang, "Self-attention mechanism based dynamic fault diagnosis and classification for chemical processes," *J. Phys., Conf.*, vol. 1914, no. 1, May 2021, Art. no. 012046.

[166] Y. Huang, J. Zhang, R. Liu, and S. Zhao, "Improving accuracy and interpretability of CNN-based fault diagnosis through an attention mechanism," *Processes*, vol. 11, no. 11, p. 3233, Nov. 2023.

[167] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[168] J. Grezmak, P. Wang, C. Sun, and R. X. Gao, "Explainable convolutional neural network for gearbox fault diagnosis," *Proc. CIRP*, vol. 80, pp. 476–481, Jan. 2019.

[169] J. Grezmak, J. Zhang, P. Wang, K. A. Loparo, and R. X. Gao, "Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3172–3181, Mar. 2020.

[170] J. Grezmak, J. Zhang, P. Wang, and R. X. Gao, "Multi-stream convolutional neural network-based fault diagnosis for variable frequency drives in sustainable manufacturing systems," *Proc. Manuf.*, vol. 43, pp. 511–518, Jan. 2020.

[171] I. Kim, S. Wook Kim, J. Kim, H. Huh, I. Jeong, T. Choi, J. Kim, and S. Lee, "Single domain generalizable and physically interpretable bearing fault diagnosis for unseen working conditions," *Exp. Syst. Appl.*, vol. 241, May 2024, Art. no. 122455.

[172] J.-H. Han, S.-U. Park, and S.-K. Hong, "A study on the effectiveness of current data in motor mechanical fault diagnosis using XAI," *J. Electr. Eng. Technol.*, vol. 17, pp. 3329–3335, Aug. 2022.

[173] N. Herwig, Z. Peng, and P. Borghesani, "Bridging the trust gap: Evaluating feature relevance in neural network-based gear wear mechanism analysis with explainable AI," *Tribol. Int.*, vol. 187, Sep. 2023, Art. no. 108670.

[174] X. Nie and G. Xie, "A novel normalized recurrent neural network for fault diagnosis with noisy labels," *J. Intell. Manuf.*, vol. 32, no. 5, pp. 1271–1288, Jun. 2021.

[175] M. Parziale, Y. F. Yeung, K. Youcef-Toumi, M. Giglio, and F. Cadini. (Nov. 2023). *Anomaly Characterization for the Condition Monitoring of Rotating Shafts Exploiting Data Fusion and Explainable Convolutional Neural Networks*. Rochester, NY, USA. [Online]. Available: https://papers.ssrn.com/abstract=4634978

[176] Z. Pan, Y. Wang, K. Wang, G. Ran, H. Chen, and W. Gui, "Layer-wise contribution-filtered propagation for deep learning-based fault isolation," *Int. J. Robust Nonlinear Control*, vol. 32, no. 17, pp. 9120–9138, Nov. 2022.

[177] H. Wu, A. Huang, and J. W. Sutherland, "Layer-wise relevance propagation for interpreting LSTM-RNN decisions in predictive maintenance," *Int. J. Adv. Manuf. Technol.*, vol. 118, nos. 3–4, pp. 963–978, Jan. 2022.

[178] D. Solís-Martín, J. Galán-Páez, and J. Borrego-Díaz, "On the soundness of XAI in prognostics and health management (PHM)," *Information*, vol. 14, no. 5, p. 256, Apr. 2023.

[179] P. Agarwal, M. Tamer, and H. Budman, "Explainability: Relevance based dynamic deep learning algorithm for fault detection and diagnosis in chemical processes," *Comput. Chem. Eng.*, vol. 154, Nov. 2021, Art. no. 107467.

[180] L. Ye, H. Wu, Y. Chen, and Z. Fei, "Interpret what a convolutional neural network learns for fault detection and diagnosis in process systems," *J. Process Control*, vol. 131, Nov. 2023, Art. no. 103086.

[181] S. Wang, Q. Zhao, Y. Han, and J. Wang, "Root cause diagnosis for process faults based on multisensor time-series causality discovery," *J. Process Control*, vol. 122, pp. 27–40, Feb. 2023.

[182] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[183] D. Wang, Z. Peng, and L. Xi, "The sum of weighted normalized square envelope: A unified framework for kurtosis, negative entropy, Gini index and smoothness index for machine health monitoring," *Mech. Syst. Signal Process.*, vol. 140, Jun. 2020, Art. no. 106725.

[184] B. Hou, D. Wang, Y. Chen, H. Wang, Z. Peng, and K.-L. Tsui, "Interpretable online updated weights: Optimized square envelope spectrum for machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 169, Apr. 2022, Art. no. 108779.

[185] J. Ding, Y. Wang, Y. Qin, and B. Tang, "Deep time–frequency learning for interpretable weak signal enhancement of rotating machineries," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106598.

[186] R. N. A. Algburi, H. Gao, and Z. Al-Huda, "A new synergy of singular spectrum analysis with a conscious algorithm to detect faults in industrial robotics," *Neural Comput. Appl.*, vol. 34, no. 10, pp. 7565–7580, May 2022.

[187] C. Yang, H. Li, and S. Cao, "Unknown fault diagnosis of planetary gearbox based on optimal rank nonnegative matrix factorization and improved stochastic resonance of bistable system," *Nonlinear Dyn.*, vol. 111, no. 1, pp. 217–242, Jan. 2023.

[188] J. Sipple, "Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9016–9025.

[189] P. Peng, Y. Zhang, H. Wang, and H. Zhang, "Towards robust and understandable fault detection and diagnosis using denoising sparse autoencoder and smooth integrated gradients," *ISA Trans.*, vol. 125, pp. 371–383, Jun. 2022.

[190] J. Du, X. Li, Y. Gao, and L. Gao, "Integrated gradient-based continuous wavelet transform for bearing fault diagnosis," *Sensors*, vol. 22, no. 22, p. 8760, Nov. 2022.

[191] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[192] M. G. Alfarizi, J. Vatn, and S. Yin, "An extreme gradient boosting aided fault diagnosis approach: A case study of fuse test bench," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 661–668, Aug. 2023.

[193] D. Zhou, Q. Yao, H. Wu, S. Ma, and H. Zhang, "Fault diagnosis of gas turbine based on partly interpretable convolutional neural networks," *Energy*, vol. 200, Jun. 2020, Art. no. 117467.

[194] B. An, S. Wang, Z. Zhao, F. Qin, R. Yan, and X. Chen, "Interpretable neural network via algorithm unrolling for mechanical fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[195] G. Li, L. Chen, C. Fan, J. Gao, C. Xu, and X. Fang, "Improved convolutional neural network chiller early fault diagnosis by gradient-based feature-level model interpretation and feature learning," *Appl. Thermal Eng.*, vol. 236, Jan. 2024, Art. no. 121549.

[196] A. Malhi, M. Madhikermi, M. Huotari, and K. Främling, "Air handling unit explainability using contextual importance and utility," in *Proc. Int. Conf. Mobile Ubiquitous Syst., Comput., Netw., Services*. New York, NY, USA: Springer, 2021, pp. 513–519.

[197] D. F. N. Oliveira, L. F. Vismari, A. M. Nascimento, J. R. de Almeida, P. S. Cugnasca, J. B. Camargo, L. Almeida, R. Gripp, and M. Neves, "A new interpretable unsupervised anomaly detection method based on residual explanation," *IEEE Access*, vol. 10, pp. 1401–1409, 2022.

[198] Y. Zhuo, J. Qian, Z. Song, and Z. Ge, "ABIGX: A unified framework for eXplainable fault detection and classification," 2023, *arXiv:2311.05316*.

[199] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," 2019, *arXiv:1907.09294*.

[200] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[201] S. Vollert, M. Atzmueller, and A. Theissler, "Interpretable machine learning: A brief survey from the predictive maintenance perspective," in *Proc. 26th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2021, pp. 01–08.

[202] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[203] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[204] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 210–215.

[205] T. Yan, D. Wang, J. Kong, and Z. Peng, "Large margin-learning methodology from time-frequency maps and its physically interpretable weights for simultaneous machine health monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 200, Oct. 2023, Art. no. 110615.

[206] R. R. A. Harinarayan and S. M. Shalinie, "XFDDC: EXplainable fault detection diagnosis and correction framework for chemical process systems," *Process Saf. Environ. Protection*, vol. 165, pp. 463–474, Sep. 2022.

[207] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019.

[208] F. Rao, M. Zeng, and Y. Cheng, "A novel interpretable model via algorithm unrolling for intelligent fault diagnosis of machinery," *IEEE Sensors J.*, vol. 24, no. 1, pp. 495–505, Jan. 2024.

[209] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, and R. X. Gao, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 4, pp. 2302–2312, Apr. 2022.

[210] Z. Wang, J. Xuan, and T. Shi, "Multi-source information fusion deep self-attention reinforcement learning framework for multi-label compound fault recognition," *Mechanism Mach. Theory*, vol. 179, Jan. 2023, Art. no. 105090.

[211] G. Jiang, C. Jia, S. Nie, X. Wu, Q. He, and P. Xie, "Multiview enhanced fault diagnosis for wind turbine gearbox bearings with fusion of vibration and current signals," *Measurement*, vol. 196, Jun. 2022, Art. no. 111159.

[212] J. Tang, J. Wu, B. Hu, and J. Liu, "An intelligent diagnosis method using fault feature regions for untrained compound faults of rolling bearings," *Measurement*, vol. 204, Nov. 2022, Art. no. 112100.

[213] G. Jiang, J. Wang, L. Wang, P. Xie, Y. Li, and X. Li, "An interpretable convolutional neural network with multi-wavelet kernel fusion for intelligent fault diagnosis," *J. Manuf. Syst.*, vol. 70, pp. 18–30, Oct. 2023.

[214] M. Sadoughi and C. Hu, "Physics-based convolutional neural network for fault diagnosis of rolling element bearings," *IEEE Sensors J.*, vol. 19, no. 11, pp. 4181–4192, Jun. 2019.

[215] Q. Wu, X. Ding, L. Zhao, R. Liu, Q. He, and Y. Shao, "An interpretable multiplication-convolution sparse network for equipment intelligent diagnosis in antialiasing and regularization constraint," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[216] Q. Chen, X. Dong, G. Tu, D. Wang, C. Cheng, B. Zhao, and Z. Peng, "TFN: An interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis," *Mech. Syst. Signal Process.*, vol. 207, Jan. 2024, Art. no. 110952.

[217] Z. Shang, Z. Zhao, and R. Yan, "Denoising fault-aware wavelet network: A signal processing informed neural network for fault diagnosis," *Chin. J. Mech. Eng.*, vol. 36, no. 1, p. 9, Jan. 2023.

[218] M. Gwak, M. S. Kim, J. P. Yun, and P. Park, "Robust and explainable fault diagnosis with power-perturbation-based decision boundary analysis of deep learning models," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 6982–6992, May 2023.

[219] A. K. M. Nor, S. R. Pedapati, M. Muhammad, and V. Leiva, "Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data," *Mathematics*, vol. 10, no. 4, p. 554, Feb. 2022.

[220] W.-T. Yang, M. S. Reis, V. Borodin, M. Juge, and A. Roussy, "An interpretable unsupervised Bayesian network model for fault detection and diagnosis," *Control Eng. Pract.*, vol. 127, Oct. 2022, Art. no. 105304.

[221] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[222] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Jun. 2010, pp. 399–406.

[223] Q. Qian, F. Xiong, and J. Zhou, "Deep unfolded iterative shrinkage-thresholding model for hyperspectral unmixing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 2151–2154.

[224] J. Zhang and B. Ghanem, "ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1828–1837.

[225] R. Khatib, D. Simon, and M. Elad, "Learned greedy method (LGM): A novel neural architecture for sparse coding and beyond," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103095.

[226] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 521–538, Mar. 2020.

[227] Z. Feng, Y. Zhou, M. J. Zuo, F. Chu, and X. Chen, "Atomic decomposition and sparse representation for complex signal analysis in machinery fault diagnosis: A review with examples," *Measurement*, vol. 103, pp. 106–132, Jun. 2017.

[228] Z. Zhao, T. Li, B. An, S. Wang, B. Ding, R. Yan, and X. Chen, "Model-driven deep unrolling: Towards interpretable deep learning against noise attacks for intelligent fault diagnosis," *ISA Trans.*, vol. 129, pp. 644–662, Oct. 2022.

[229] F. Qin, S. Wang, S. Wang, Z. Zhao, R. Yan, and X. Chen, "MCAN: Interpretable multi-scale component analysis network for mechanical fault diagnosis," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2023, pp. 1–6.

[230] Z. Kong, A. Jones, and C. Belta, "Temporal logics for learning and detection of anomalous behavior," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1210–1222, Mar. 2017.

[231] G. Chen, Y. Lu, R. Su, and Z. Kong, "Interpretable fault diagnosis of rolling element bearings with temporal logic neural network," 2022, *arXiv:2204.07579*.

[232] R. Tian, M. Cui, and G. Chen, "A neural-symbolic network for interpretable fault diagnosis of rolling element bearings based on temporal logic," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[233] G. Chen, P. Wei, H. Jiang, and M. Liu, "Formal language generation for fault diagnosis with spectral logic via adversarial training," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 119–129, Jan. 2022.

[234] G. Chen, M. Liu, and J. Chen, "Frequency-temporal-logic-based bearing fault diagnosis and fault interpretation using Bayesian optimization with Bayesian neural networks," *Mech. Syst. Signal Process.*, vol. 145, Nov. 2020, Art. no. 106951.

[235] G. Chen, Y. Lu, and R. Su, "Interpretable fault diagnosis with shapelet temporal logic: Theory and application," *Automatica*, vol. 142, Aug. 2022, Art. no. 110350.

[236] T. Yan, D. Wang, and Y. Wang, "Discrimination- and sparsity-driven weight-oriented optimization model for interpretable initial fault detection and fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[237] H. Pu, K. Zhang, and Y. An, "Restricted sparse networks for rolling bearing fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 11139–11149, Nov. 2023.

[238] L. Ma, Y. Ding, Z. Wang, C. Wang, J. Ma, and C. Lu, "An interpretable data augmentation scheme for machine fault diagnosis based on a sparsity-constrained generative adversarial network," *Exp. Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115234.

[239] J. Yuan, S. Cao, G. Ren, F. Su, H. Jiang, and Q. Zhao, "LW-net: An interpretable network with smart lifting wavelet kernel for mechanical feature extraction and fault diagnosis," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 15661–15672, Sep. 2022.

[240] S. Li, T. Li, C. Sun, X. Chen, and R. Yan, "WPConvNet: An interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 15, 2023, doi: 10.1109/TNNLS.2023.3282599.

[241] W. Cheng, X. Liu, J. Xing, X. Chen, B. Ding, R. Zhang, K. Zhou, and Q. Huang, "AFARN: Domain adaptation for intelligent cross-domain bearing fault diagnosis in nuclear circulating water pump," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 3229–3239, Mar. 2023.

[242] F. B. Abid, M. Sallem, and A. Braham, "Robust interpretable deep learning for intelligent fault diagnosis of induction motors," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3506–3515, Jun. 2020.

[243] M. Kraus and S. Feuerriegel, "Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences," *Decis. Support Syst.*, vol. 125, Oct. 2019, Art. no. 113100.

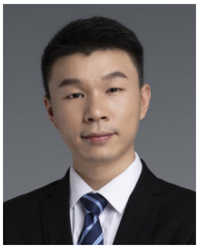
[244] G. Hajgató, R. Wéber, B. Szilágyi, B. Tóthpál, B. Gyires-Tóth, and C. Hos, "PredMaX: Predictive maintenance with explainable deep convolutional autoencoders," *Adv. Eng. Informat.*, vol. 54, Oct. 2022, Art. no. 101778.

GANG CHEN (Member, IEEE) received the bachelor's and master's degrees in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012 and 2015, respectively, and the Ph.D. degree in mechanical and aerospace engineering from the University of California at Davis, Davis, CA, USA, in 2020. From 2020 to 2021, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His research interests include machine learning, formal methods, control, signal processing, and fault diagnosis.

JUNLIN YUAN received the B.S. degree in mechanical engineering from Wuhan Institute of Technology University, in 2019, and the M.S. degree in mechanical electronics from Shantou University, in 2023, with a research focus on deep learning and in-suit quality monitoring in additive manufacturing. He is currently pursuing the Ph.D. degree in mechanical engineering with South China University of Technology. His current research interest includes interpretable intelligent fault diagnosis.

YIYUE ZHANG received the B.S. degree in intelligence manufacturing from Hefei University of Technology, Anhui, China, in 2023. He is currently pursuing the master's degree with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology. His research interests include physical-informed neural networks, interpretable intelligent fault diagnosis, and neural network applications.

HANYUE ZHU received the bachelor's degree in automation from Sichuan University, Sichuan, China, in 2023. She is currently pursuing the master's degree with the Shien-Ming WU School of Intelligent Engineering of Control Science and Engineering, South China University of Technology. Her current research interest includes interpretable intelligent fault diagnosis.



RUYI HUANG (Member, IEEE) received the Ph.D. degree in mechanical engineering from South China University of Technology (SCUT), Guangzhou, China, in 2021. He is currently a Postdoctoral Researcher with the Shien-Ming Wu School of Intelligent Engineering, SCUT. His research interests include intelligent fault diagnosis (IFD), prognostics and health management (PHM), and multisensory data/information fusion technology for condition-based maintenance. He was awarded the Graduate Fellowship Award by the IEEE Instrumentation and Measurement Society in 2020 and the Best Paper Award by ICSMD 2023. He serves as an Associate Editor for IEEE Open Journal of Instrumentation and Measurement and an Active Peer Reviewer for several international journals, such as IEEE Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, and IEEE Transactions on Instrumentation and Measurement. He was recognized as the Outstanding Reviewer for IEEE TIM in 2021 and 2022.

FENGTAO WANG received the M.S. degree in mechanical engineering from Jilin University of Technology, Jilin, China, in 2000, and the Ph.D. degree in mechanical engineering from Dalian University of Technology, Dalian, China, in 2003. He is currently a Professor of Department of Mechanical Engineering, College of Engineering, Shantou University, Shantou, China. His current research interests include signal processing, machine learning, metal-based additive manufacturing process monitoring and defect identification, and rotating machinery fault diagnosis. He has received the Second Prize of the National Scientific and Technological Progress Award as the first accomplisher of the Dalian University of Technology.

WEIHUA LI (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003. He is currently the Dean and a Professor with the School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, China. His research interests include industrial intelligence, industrial big data, digital twins, intelligent maintenance and health management, and intelligent connected vehicles. He is serving as the Co-Chair for the Technical Committee (TC-3) on Condition Monitoring and Fault Diagnosis Instrument, IEEE Instrumentation and Measurement Society (IM Society). He serves as a member of the editorial board of IEEE Transactions on Instrumentation and Measurement, IEEE Sensors Journal, *Chinese Journal of Mechanical Engineering*, Journal of Dynamics, Monitoring and Diagnostic, and Journal of Vibration Engineering, in Chinese.