scientific reports

Check for updates

OPEN

Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments

Ali Saeed[1,4]✉, Muazzam A. Khan[1,4], Usman Akram[2], Waeal J. Obidallah[3]✉, Soyiba Jawed[2] & Awais Ahmad[3]

Industry 4.0 represents the fourth industrial revolution, which is characterized by the incorporation of digital technologies, the Internet of Things (IoT), artificial intelligence, big data, and other advanced technologies into industrial processes. Industrial Machinery Health Management (IMHM) is a crucial element, based on the Industrial Internet of Things (IIoT), which focuses on monitoring the health and condition of industrial machinery. The academic community has focused on various aspects of IMHM, such as prognostic maintenance, condition monitoring, estimation of remaining useful life (RUL), intelligent fault diagnosis (IFD), and architectures based on edge computing. Each of these categories holds its own significance in the context of industrial processes. In this survey, we specifically examine the research on RUL prediction, edge-based architectures, and intelligent fault diagnosis, with a primary focus on the domain of intelligent fault diagnosis. The importance of IFD methods in ensuring the smooth execution of industrial processes has become increasingly evident. However, most methods are formulated under the assumption of complete, balanced, and abundant data, which often does not align with real-world engineering scenarios. The difficulties linked to these classifications of IMHM have received noteworthy attention from the research community, leading to a substantial number of published papers on the topic. While there are existing comprehensive reviews that address major challenges and limitations in this field, there is still a gap in thoroughly investigating research perspectives across RUL prediction, edge-based architectures, and complete intelligent fault diagnosis processes. To fill this gap, we undertake a comprehensive survey that reviews and discusses research achievements in this domain, specifically focusing on IFD. Initially, we classify the existing IFD methods into three distinct perspectives: the method of processing data, which aims to optimize inputs for the intelligent fault diagnosis model and mitigate limitations in the training sample set; the method of constructing the model, which involves designing the structure and features of the model to enhance its resilience to challenges; and the method of optimizing training, which focuses on refining the training process for intelligent fault diagnosis models and emphasizes the importance of ideal data in the training process. Subsequently, the survey covers techniques related to RUL prediction and edge-cloud architectures for resource-constrained environments. Finally, this survey consolidates the outlook on relevant issues in IMHM, explores potential solutions, and offers practical recommendations for further consideration.

Keywords Intelligent fault diagnosis, Deep learning, Transfer learning, Industrial Internet of Things, Edge computing, Remaining useful life

[1]Department of Computer Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan. [2]Department of Computer and Software Engineering, NUST College of Electrical and Mechanical Engineering, Islamabad 44000, Pakistan. [3]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11673, Saudi Arabia. [4]ICESCO Chair Big Data Analytics and Edge Computing, Quaid-i-Azam University, Islamabad 45320, Pakistan. ✉email: alisaeed101@gmail.com; wobaidallah@imamu.edu.sa

Industrial machinery health management (IMHM) is a comprehensive approach to monitor, assess, and optimize the condition and performance of machinery used in industrial settings. Industrial machinery is the equipment found in different industrial setups e.g. manufacturing, packaging, assembling, and quality assurance. These mechanical assets include fans, motors, and pumps, which are prone to wear and tear and are monitored for fault detection and life prediction. The condition of a machine is assessed based on the data gathered over the service period. A correct and well-timed assessment can help the maintenance team to take proactive measures and avoid failures[1]. The increasing complexity of modern industrial systems, machinery, and technologies has made it challenging to manually monitor and diagnose faults effectively[2]. Organizations have shifted from reactive and time-based maintenance approaches to proactive strategies that prevent unplanned downtime[3]. Unplanned downtime and unexpected failures can result in significant financial losses for industries. These proactive strategies are heavily reliant on data-driven approaches[4]. One of how data-driven approaches contribute to proactive strategies is intelligent fault diagnosis (IFD). IFD relies on data collected from systems and machinery to detect anomalies and diagnose faults by analyzing data patterns. These data patterns include time-series, multivariate, spatial, categorical, event-based, and frequency domain data[5]. IFD often involves extracting features from the raw data and analyzing patterns in a feature space. In IMHM, time-series data is one of the most prevalent types[6]. It involves collecting measurements over time to monitor the behavior of industrial assets. Typical measures include sensor readings such as temperature, vibration, and pressure collected at regular intervals.

The research community has focused on intelligent fault diagnosis in recent years based on the advancements in sensor technology, communication, and artificial intelligence (AI). Cumulative research has conclusively shown that the synergy between extensive datasets and advanced AI technology not only significantly enhances the precision and reliability of intelligent fault diagnosis but also effectively mitigates the inherent challenges associated with fault prediction and identification[7–13]. IFD methods are classified into three categories, Data processing, model creation, and training optimization. Data processing methods include approaches such as signal processing, feature extraction, and data normalization[6]. As IFD derives diagnostic insights from data, the availability, distribution, and preprocessing of the data can directly impact the effectiveness of fault diagnosis[14]. Model creation methods in IFD focus on developing algorithms and models capable of detecting and diagnosing faults based on the processed data. Optimization methods focus on hyperparameters optimization, model-architecture fine-tuning, and addressing issues like overfitting and underfitting. Researchers have focused on each category to improve the IFD methods and their effectiveness[15,16]. Data processing methods can be further categorized with respect to published research in the context of IFD. These categories are signal processing, feature extraction, feature selection, normalization, scaling, dimensionality reduction, data imputation, cleaning, fusion, time-synchronization, and balancing. Feature extraction and feature selection, though related, involve fundamentally different approaches. Feature extraction creates new features from the raw data by transforming it into a lower-dimensional space, capturing the most informative aspects of the data, often using methods such as Principal Component Analysis (PCA) or autoencoders[17]. On the other hand, feature selection involves choosing a subset of the most relevant features from the existing set without altering the original data structure, based on criteria like mutual information or correlation with the target variable[18]. Data processing techniques go way back than artificial intelligence and there has been a lot of research in this category in the last two decades. The data processing methods have evolved with time as needed by the processing techniques. Each category has a substantial number of publications in recent years which we have tried to include in the review. Signal processing techniques manipulate, analyze and transform signals, such as Fourier transforms and wavelet analysis[19–24]. Feature extraction involves identifying and selecting relevant features or characteristics from raw data that are informative for a particular task. Yang et al.[21] have proposed a method based on variational mode decomposition (VMD) and improved envelope spectrum entropy (IESE). Enhancing classification accuracy and reducing diagnosis time, feature selection involves the identification of optimal features. The feature selection method proposed in Ref.[19] chooses optimal features between classes to obtain a feature ranking that is conducive to classification. Chun-yao lee et al. have put forward a feature selection method based on based on memory space computation genetic algorithm[20]. A hybrid approach for diagnosing faults in rolling bearings leveraging continuous wavelet transform (CWT), convolutional neural network (CNN), and support vector machine (SVM) is presented in Ref.[22]. IFD in rotating machinery faces the problem of inconsistency in feature distribution and lack of labeled fault feature data[25]. A lot of research has been directed in the field of domain adaptation (DA) where two of its techniques are majorly studied i.e. Adversarial training and transfer learning. Xuejun Liu et al. have proposed a generative adversarial network that can handle imbalanced fault samples, the framework captures distribution information by fusing two pieces of information from two domains[26]. Huo et al. has used a self-attention mechanism with residual networks to extract features GAN fitting the data distribution to generate high-quality samples of balanced dataset[27]. This survey will comprehensively review and discuss the research advancements aimed at addressing the data processing challenges.

Model creation methods in the context of IFD involve developing algorithms and models capable of detecting and diagnosing faults based on processed data. These categories are supervised, unsupervised, hybrid models, ensemble learning, deep learning, rule-based systems, time-series analysis models, transfer learning, and meta-learning models. In ensemble learning techniques, Tong et al.[28] have come up with an ensemble learning-based multi-sensor information fusion method. Ensemble EMD is used in combination with fuzzy rough sets (FRS) by Zhou et al. to remove noise from acquired data and enhance the features[29]. Xiong et al. has proposed a method that reconstructs indicators from input data based on ensemble empirical mode decomposition and maps these indicators to gramian angular fields (GAF)[30]. Yu et al. have proposed a framework that uses ResNet with multiscale attention mapping to extract relevant features[31] and ensemble EMD is used as a first step to process the data for improving the neural networks extraction ability. In recent years, transfer learning has garnered significant attention from research communities and is acknowledged for its ability to establish connections

between additional testing and training samples, leading to expedited output and efficient results[32]. Wenying et al. has proposed a method that fuses adversarial learning and means discrepancy[33]. In another paper[34] same group of authors have proposed a novel transfer learning method that uses different distributed domains to transfer knowledge. A novel deep transfer diagnosis model for bearing faults, based on the Residual Network (ResNet), has been proposed by Yu et al.[35].

An important and well-researched category in model development focuses on analyzing time-series data and capturing temporal dependencies. The bulk of papers have combined some base analysis techniques like long short-term memory (LSTM) with convolution neural networks. Zheng et al. have proposed an end-to-end deep learning method by applying one-dimensional CNN and LSTM[8]. Abiodun et al. presented a causal augmented convolution network with implementation for long-sequence time series prediction[36]. Shen et al. developed a transformer for analyzing time series data even under sharp variation[37].

Once the model is built its efficiency is as good as the training it gets. Training optimization methods in the context of IFD involve improving the efficiency and effectiveness of the training process for models. Researchers have focused on several categories of optimization methods, which include hyperparameter tuning, regularization, learning rate schedules, data augmentation, ensemble methods, quantization, and transfer learning. Insufficient fault data poses a challenge, given the inherent difficulty in acquiring authentic industrial data from machinery operating under fault conditions[38]. Using multiple sensors to learn fault data may lead to more accurate diagnostic results by fusing information from these multiple sources. The performance of deep learning-based methods for diagnosing rolling bearing faults is substantially compromised due to the scarcity of operating data arising from intricate and variable working conditions. To handle this issue Gao et al. have augmented data by using generative adversarial network (GAN)[39]. Luo et al. has used GAN meta-learning to achieve a higher accuracy[40]. To ensure similarity among real and generated samples and improved diversity of the generated samples Ren et al. have used multiscale feature fusion with GAN[27,41]. Lu et al. has used GAN to generate data closer to the target domain supplemented by inverse attention to capture the learning potential of all features[42]. Another particular facet within the realm of IFD is the prediction of Remaining Useful Life (RUL), a topic that has garnered significant attention in the research community. Ensuring the quality of manufactured or assembled products relies on effective condition-based monitoring (CbM) of industrial equipment[43]. However, there is a growing emphasis on estimating and predicting the remaining useful life (RUL) of the devices, such as machines and robots, constituting the production or assembly line[44]. Mao et al. has proposed a deep learning-based approach to predict the useful life of bearings[45]. Spectral correlation establishes a health indicator (HI) that is used later to get rid of any fluctuations in life prediction[46]. Hui Gao et al. have explored both dimensions of data (i.e. time and space) collected from multiple sensors and predicted the life of bearings using DL dual channel feature attention and Bi-directional long short-term memory (Bi-LSTM) for feature extraction[47]. Smart manufacturing leverages the Industrial Internet of Things (IIoT) for IFD, integrating IoT technologies, Big Data techniques, artificial intelligence, cloud computing, and other continuously evolving enabling technologies[48]. Zhao et al. proposed a framework that combines compressed sensing and edge computing (EC) into a wireless sensor network (WSN)[49]. TinyML holds the promise of being a significant facilitator for smart sensing nodes in the IFD of machines. This is achieved by integrating robust machine learning algorithms into low-cost edge devices[50].

Monitoring the industrial components in real-time require setup of resource constrained devices i.e. computing devices with limited data acquisition, storage and processing capabilities[51]. Resource-constrained environments are limited in terms of processing power, memory space, energy consumption and network capability with connectivity. These limitations can cause hindrance in carrying out complex computing tasks e.g. analyzing historic data for pattern detection[52]. Edge computing architectures, embedded systems, IoT devices, wireless sensor networks and industrial control systems, combination of these technologies are used in modern day industrial setups which play a significant and critical role in maintaining these setups.

In contrast to prior comprehensive literature reviews, this survey makes distinctive contributions in the following ways:

1. This research delves into the prevailing research challenges and limitations within the domains of IFD, RUL prediction, and edge-based architectures.
2. This survey focuses on IFD from the perspectives of data preprocessing, techniques, and approaches. The entire developmental journey of IFD is comprehensively examined.
3. This paper furnishes practical instances of IFD challenges within the realm of fault diagnosis, shedding light on aspects such as data availability and on-machine diagnosis-areas that have garnered limited attention in existing research.
4. This survey addresses the tangible limitations and challenges associated with the real-time implementation of IMHM utilizing the Industrial Internet of Things (IIoT).
5. Through an in-depth analysis of a multitude of research works, this survey consolidates the limitations observed in existing studies and outlines future research trends. It offers a systematic perspective for seasoned researchers and serves as a fundamental tutorial for those new to the field.The remaining part of the article is structured as follows. "Research methodology and initial analysis" section outlines the adopted research methodology. Publicly available benchmarks and the most widely used datasets in the field for rolling bearing fault diagnosis are discussed in "Datasets overview" section. Intelligent fault diagnosis methods and their implementation challenges are discussed in "Intelligent fault diagnosis" section. The role of remaining useful life (RUL) prediction in IMHM is covered in "Role of remaining useful life (RUL) prediction in IMHM" section. For real-time predictions and resource restraint environments cloud-edge architectures are discussed with their applications and implementation challenges "Cloud-edge enabled real-time fault diagnosis in

Sr.	Source
1.	Google Scholar
2.	IEEE Explore
3.	Science Direct
4.	ACM Digital Library

Table 1. Research sources for exploration.

Category	Perspective	Methods	References	Year(s)
Intelligent fault diagnosis (IFD)	Data processing	Feature extraction	54–75	2023
		Data augmentation	40,76–84	2023
		Imbalanced learning	85–89	2023
		Noise removal	24,66,90–93	2023
		Multi-scale time series analysis	45,94–104	2023
	Model creation	Neural networks	55,73,105–117	2023
		Convolution neural networks	56,106–109,111–113,116–124	2023
		Recurrent neural networks	125–132	2023
		LSTM methods	133–136	2023
		Transformer-based methods	137–140	2023
		Auto encoders	141–144	2023
		Ensemble learning	105,145–147	2023
	Training optimization	Hyperparameter tuning	148–151	2023
		Regularization techniques	152–154	2023
		Incremental learning	124,129–131	2023
		Data augmentation	79,119,122,155,156	2023
		Transfer learning	57,72,121,139,157–170	2023
Remaining useful life (RUL)			58,94,95,171–175	2023
Cloud-edge architecture		Genetic algorithm (GA), NNs, LSTM	50,60,68,176–189	2023

Table 2. A summarized overview of related work in the field of intelligent industrial machinery health management and fault diagnosis.

IMHM" section. The survey is concluded in "Conclusion and future prospective" section highlighting future perspectives.

Research methodology and initial analysis

Research methodology

This survey systematically explored publications in the field of intelligent fault diagnosis. The recent research is explored the three main categories of IFD, extending to work related to the classification of faults in real-time and the prediction of the remaining useful life of rotating bearings. It comprehensively covered various aspects, concluding with an examination of its implementation and challenges within the edge-cloud infrastructure. The investigation encompassed sources from prominent research exploration platforms i.e. Google Scholar, Elsevier, Springer, IEEE Explore, and ACM Digital, summarized in Table 1. Relevant publications from all databases and across all publication dates (up to November 2023) were compiled for analysis. Two categories of keywords were employed in the search. The first set, 'Predictive Health Maintenance', was utilized to define the research area, while the second set, 'intelligent fault diagnosis,' was employed to specify the research topic. Regarding fault diagnosis, additional qualifiers were applied, including 'bearing faults', 'industrial machinery', 'data-driven approaches', and 'Rolling-bearing fault diagnosis'. In the context of real-world application, we incorporated qualifiers such as 'intelligent framework', 'edge-cloud computing', and 'PHM applications'.

Initial analysis

In the start of the search process, a total of 300 publications were shortlisted. Upon additional in-depth review, 194 publications relevant to the survey scope were selected. To keep the survey related to most recent research developments the timeline is restricted to 2020 to 2023, summarized in Table 2. The domain of Intelligent fault diagnosis got the spotlight after 2016 followed by the prevalence of deep learning and convolution neural networks[53]. Intelligent Fault Diagnosis of rotatory machinery has more significance as the industrial revolution is enabling industries to embed IoT in their daily operations and processes leading to a huge amount of machine health-related data. Research in the last 5 years has focused on the use of deep learning techniques and implementation of these techniques in the most efficient way on the hardware for real-time maintenance decision support systems. The published research is explored categorically to excavate relevant domain knowledge in

each category with respect to implementation and optimization. Figure 1 depicts the detailed categories and techniques in the field of intelligent fault diagnosis.

Classification of intelligent fault diagnosis (IFD) methods

In the domain of industrial machinery health management (IMHM), assets are monitored and data is collected based on which the predictive maintenance (PdM) system recommends actions. In this domain, Intelligent fault diagnosis refers to the use of advanced technologies, particularly artificial intelligence (AI) and machine learning (ML), to detect and diagnose faults or abnormalities in systems. This can be applied across various domains such as manufacturing, healthcare, telecommunications, and more. The primary goal is to enhance the accuracy and efficiency of fault detection and diagnosis processes. In many complex systems, the occurrence of faults or anomalies can lead to system failures, reduced efficiency, and increased maintenance costs. Traditional fault diagnosis methods often rely on predefined rules or manual analysis, which may not be effective in handling complex and dynamic systems. The key challenges in this domain are:

- Modern systems are becoming increasingly complex with interconnected components and dynamic behaviors. Traditional methods may struggle to cope with the intricacies of such systems[190].
- Intelligent Fault Diagnosis relies on large volumes of data from sensors, logs, and other sources. Managing and analyzing diverse data types poses a challenge, and there may be a need for effective feature extraction[34,191].
- Systems often operate in dynamic environments where normal behavior can evolve over time. Adapting to these changes and distinguishing between normal variations and actual faults is a significant challenge[192].
- Acquiring labeled training data for various fault scenarios can be resource-intensive. Ensuring that the model is trained on a representative dataset is crucial for accurate fault detection[193,194].
- Some applications, such as critical infrastructure or manufacturing processes, require real-time fault diagnosis. Achieving low-latency processing while maintaining high accuracy is a challenge[195,196].
- In many domains, especially where human safety is a concern, understanding why a system flagged a particular fault is essential. Ensuring the interpretability and explainability of intelligent fault diagnosis systems is crucial for user acceptance and trust[197].The classification of intelligent fault diagnosis methods refers to the systematic organization and categorization of approaches, techniques, or methodologies employed in identifying and diagnosing faults in complex systems using intelligent or smart technologies. Intelligent fault diagnosis methods rely on three essential tasks: health status data processing, artificial intelligence algorithms development, and diagnostic targets or optimization[53].

The key objective of data processing is to clean and prepare the raw data for analysis, ensuring it is suitable for the subsequent phases in the process of IFD. Researchers in the recent years have focused on two main challenges in application of intelligent fault diagnosis, one is insufficient fault data and second is the imbalance data[198]. There are other challenges in data processing which are studied in the literature. These challenges include data cleaning, where we handle missing values, outliers and inconsistencies in the data[43]. Next is normalization,

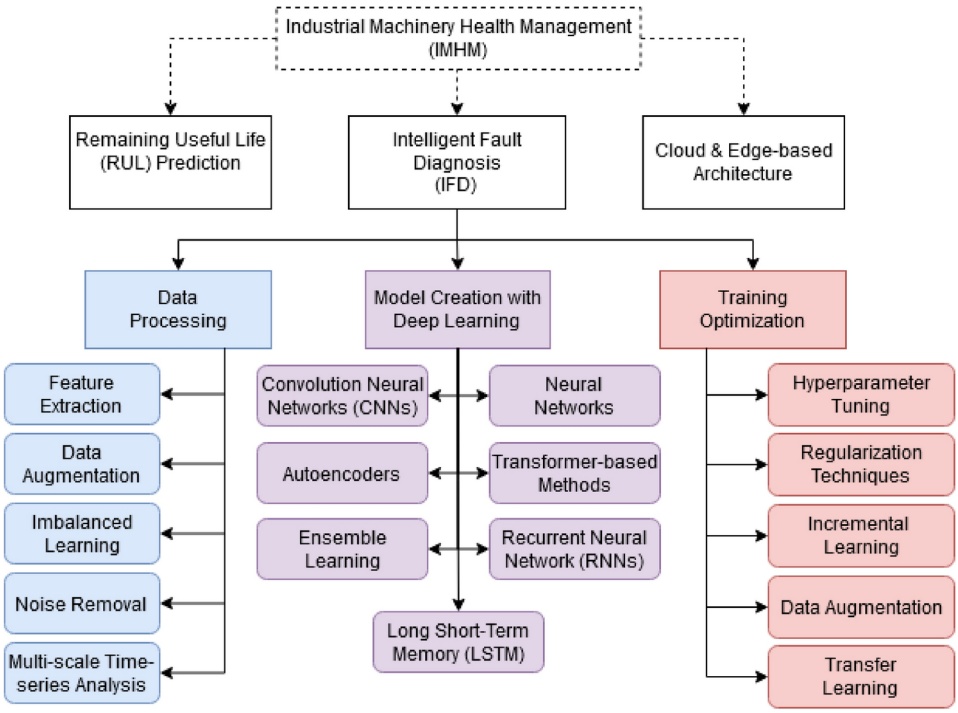

Fig. 1. Literature review scheme followed to cover all aspects of Industrial machinery health management (IMHM).

where numerical features are scaled to a standard range to ensure consistency[199]. The core of data processing is feature extraction that identify relevant feature or extract informative characteristics from the data[200]. Another main topic of interest is, data transformation i.e. convert data into a suitable format for analysis, such as time series decomposition or dimensionality reduction[201]. Sample generation is considered as a feasible solution when it comes to missing values or imbalance of the data, researchers have proposed a lot of different strategies in this domain[202].

The second phase is model development which utilizes various techniques and algorithms to analyze the processed data and identify patterns indicative of faults. This phase starts with statistical analysis and AI application i.e. application of algorithms for classification, clustering, or regression to learn patterns in the data. The next task is to train the chosen machine learning models on training data[203]. Unsupervised learning is a task focused on exploring unsupervised methods for anomaly detection or clustering in the absence of labelled fault data, this is one of the more complex tasks in the intelligent fault diagnosis process[204]. The way to improvement is assessment, hence the performance of the models is assessed in this task by using appropriate metrics (e.g., accuracy, precision, recall, F1 score)[170]. Last task in this phase is optimization i.e. fine-tuning of model parameters to enhance performance.

The last phase is making informed maintenance decisions based on the analyzed data regarding the presence or absence of faults and prediction of faults leading to failures. Initially the task is to identify the type of fault and its classification[205]. In uncertainty analysis we assess the confidence or uncertainty associated with the fault prediction. The next task is to establish decision thresholds to balance false positives and false negatives based on the system's requirements[206]. The last task of the process is to facilitate communication between the intelligent fault diagnosis system and human operators, ensuring a collaborative decision-making process[207].

Datasets overview

This section provides clear and concise overview of the datasets used in the studies included in our review. Effective performance in intelligent fault diagnosis hinges on access to substantial datasets. Publicly available data play a crucial role in this context, and numerous open datasets are currently accessible, encompassing various conditions and fault scenarios. Figure 2 illustrates the frequency with which each dataset, including the XJTU dataset, has been referenced in the studies reviewed as part of this work. These statistics reflect the usage within the scope of the selected literature, rather than the total number of times each dataset has been used by the broader research community.

The most useful data for training intelligent fault diagnosis methods is the actual data produced under the real operational conditions in the industrial machinery. The real challenge is the public availability of this data in processable format. Industries are not interested to process and put these real data sets on public domains. This issue leads to the production of laboratory based datasets. International institutions and universities have developed their testrigs to generate data that can be used to train fault diagnosis methods. Figure 3 shows a basic layout of testrig, which can be further customized according to the requirements of the setup. This experimental data can be further divided into two categories based on how the faults are introduced.

First, is to let the testrig run under normal conditions and data is recorded. After a particular number of revolutions bearings are removed and specific faults are induced by different methods like (etching, EDC, electric wire). The bearings are installed back and testrig is run again to record the faulty data. The generated dataset now contains multiple class data which may include normal data and different fault classes. The testrig is run under different speeds and loads, leading to data generated under various simulated conditions. The conditions can further tuned to represent realistic operational environments.

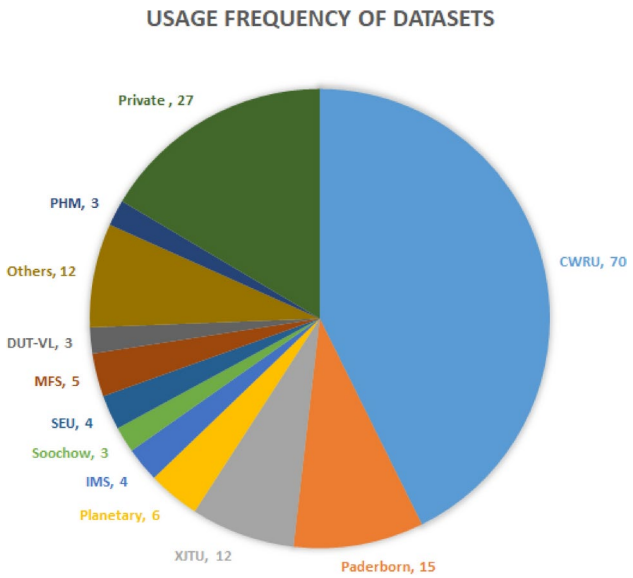

Fig. 2. Datasets utilization in literature.

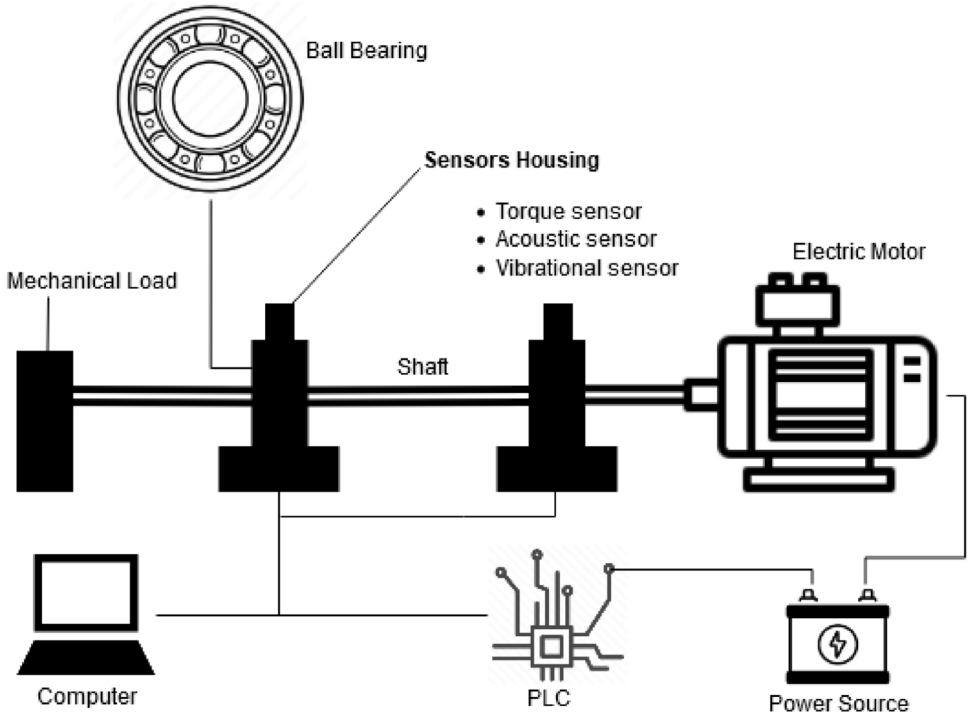

Fig. 3. Basic layout of a testrig.

Dataset	Source	Type	Data	Fault classes	Processing	Environment	Characteristics
CWRU	208	Induced faults	Time-series vibrational data	10 classes 1 × Normal 9 × Faulty	Etching	Test	Motor loads: 0–3 horsepower Motor speeds: 1797–1720 RPM Sampling Freq: 12,000 samples/s, 48,000 samples/s
Paderborn	209		Time-series vibrational and current data			Test under 4 different operating conditions	RPM: 900 & 1500 Load: 0.7, 0.1 (Nm) Radial force: 400, 1000 (N) Temp: 45–50 Celsius
XJTU	210	Run-to-Failure					
Planetary	211		Vibration Time–Frequency maps	4 Health States	S-transform		Sampling freq: 16,384 Hz
IMS bearing	212	Run-to-Failure	Time-series vibrational data				
SEU	213		Time-series vibrational data (set1) Images (set2)	1 × Normal 5 × Faulty		Test Bench 6 different conditions	System Load: 20 Hz–0 V; 30 Hz–2 V Sampling rate: 1000 samples
Soochow	214	Induced faults by wire cutting	Vibration data	13 Faults 1 × Normal 12 × Faulty	FFT normalization	Test Bench	Sample freq: 10 kHz Load: 0–1 kN Speed: 961 RPM
DUT-VL	215	Induced faults	Vibration data	10 Faults 1 × Normal 9 × Faulty		Test Bench	Sampling freq: 10 kS/s Sampling time: 30 s
PHM	216	Run-to-Failure	Vibration and temperature data			3 different conditions	1800 RPM/4000N 1650 RPM/4200N 1500 RPM/5000 N Sampling freq: 25.6 kHz (Vib) Sampling freq: 10 Hz (Temp)

Table 4. Commonly utilized datasets summary for intelligent industrial machinery health management and fault diagnosis.

In the second method, the testrig is let to run for the life of installed bearings (manufacturer specific number of revolutions) and the data is recorded till failure. The dataset generated from such setups contain 'run-to-failure' data. Table 4 lists down all the datasets used in the literature with their respective characteristics.

A ball bearing is composed of four components i.e. inner race, outer race, balls and cage laid out together and greased for smooth friction less operation, as depicted in Fig. 4. The generic industrial bearing faults are categorized based on constituent components, these faults care categorically listed in Table 3.

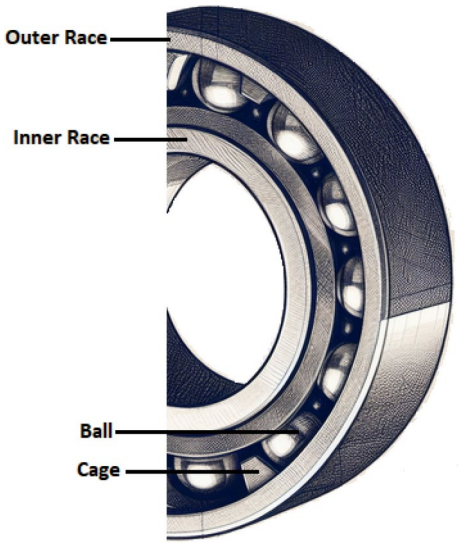

Fig. 4. Ball bearing with four basic constituent components.

Sr	Fault/failure	Severity	Fault categories
1.	Excessive preload	Moderate to high	IR, OR
2.	Brinelling	High	IR, OR
3.	Electrical damage	Moderate to high	IR, OR
4.	Runout	Moderate	IR, OR
5.	Adhesive wear	Moderate	IR, OR, Ball
6.	Fatigue failure	High	IR, OR, Ball
7.	Corrosion	Moderate to high	IR, OR, Ball
8.	Pitting	Moderate to high	IR, OR, Ball
9.	Spalling	Moderate to high	Ball
10.	Crushing	High	
11.	Surface roughness	Low to moderate	
12.	Out-of-round balls	Low to moderate	
13.	Cracking	Moderate to high	Cage
14.	Fracture	High	
15.	Wear	Low to moderate	
16.	Jamming	Moderate to high	
17.	Misalignment	Moderate	
18.	Looseness	Moderate	
19.	Ball pocket wear	Low to moderate	
20.	High-speed issues	Moderate to high	
21.	Heat damage	Moderate to high	
22.	Ball entrapment	Moderate to high	

Table 3. Ball bearing faults categorization.

The most widely used benchmark dataset in research is CWRU bearing fault dataset[208]. The Case School of Engineering at Case Western Reserve University conducted experiments using a 2 hp reliance electric motor, data was measured and carefully documented for actual test conditions and fault status. Faults are induced on motor bearings using electro-discharge machining (EDM). Defects with diameters ranging from 0.007 inches to 0.040 inches were individually introduced at the inner raceway, rolling element (i.e., ball), and outer raceway. Subsequently, the defective bearings were reinstalled into the test motor, and vibration data were recorded for motor loads varying from 0 to 3 horsepower, corresponding to motor speeds ranging from 1797 to 1720 RPM. Bearing data center from the University of Paderborn provides[209] to enable and encourage bearing condition monitoring. The testing apparatus was run under various operational scenarios to validate the effectiveness of condition monitoring methods across a range of operating conditions. Operation parameters are rotational speed, radial force, and load torque, all three parameters were kept constant for each measurement. Another

parameter is temperature, which was kept at 45 to 50 degrees centigrade during all experiments. The dataset provides a detailed description of the bearings and their respective fault categories. The data contains current and vibration data with 26 damaged bearing states and 6 undamaged (healthy) states for reference. The dataset also comes with uniform fact sheets (for systematic description of damages) and measuring log. The bearing datasets from XJTU-SY[210] were supplied by the Institute of Design Science and Basic Component at Xi'an Jiaotong University in collaboration with Changxing Sumyoung Technology Co. The XJTU-SY datasets comprised run-to-failure data for fifteen bearings across three distinct operating conditions[217]. Information was gathered at a frequency of 2.56 kHz, with 32,768 data points recorded for each sampling session. The sampling period was set at one minute. The fault elements in this dataset are the outer ring, cage, inner ring, and rolling element. The gearbox datasets from Southeast University (SEU) were furnished by the institution[213]. These SEU datasets encompass two sub-datasets, one focused on bearings and the other on gears. Both datasets were obtained using a drivetrain dynamic simulator (DDS). Two distinct operating conditions were established, each characterized by a specific rotating speed-load configuration (RS-LC): one set at 20 Hz with 0 V, and the other at 30 Hz with 2 V. IMS bearing datasets[212] are provided by the Center of intelligent maintenance systems (IMS), university of Cincinnati. The dataset contains bearing acceleration data generated by three different run-to-failure experiments on a loaded shaft. At the end of the experiments, the failure occurred in different locations of the bearings. The PHM 2012 bearing datasets[216] were employed for the PHM IEEE 2012 Data Challenge. Within these datasets, seventeen run-to-failure scenarios were included, comprising six for training and eleven for testing. The experiments covered three distinct loads, with data collected on vibration and temperature signals. The Dalian University of Technology (DUT) lab utilizes a rotating equipment test platform to simulate bearing failures and gather vibration data in various failure states of rolling bearings[215]. The sampling frequency is 10 kS/s and the sampling time is set to 30 s. Nine variations of abrasion faults, representing a specific type of normal bearing state, and a total of ten distinct bearing states were machined. These efforts targeted different locations of rotational bearing faults and varying degrees of abrasion.

There is a substantial number of research articles that have used public datasets along with private datasets to validate their proposed methods and frameworks. These private datasets are generated by laboratory experiments in a controlled environment. The test rigs are usually assembled from the common components that are generally used e.g. electric motor attached with a rotational shaft having bearings installed on it under specific loads and torque that is measured by dynamo-meter. Wei Liu et al. have used a similar test rig for recording the defect of the bearing outer ring to evaluate their method[218].

The data generated from a test environment may differ from real environment data in certain aspects, one of the main aspects is noise. Industrial machines often operate in intricate environments characterized by substantial environmental noise. Noisy data may arise from manufacturing inaccuracies, improper installation, variations in running speed, supported loads, or deficiencies in the lubrication of rolling bearings[215]. Different degrees of noise are combined with raw data to verify the method application under strong noises. Nevertheless, a notable challenge lies in the fact that the majority of these datasets originate from controlled laboratory settings, potentially diverging from the conditions encountered in real-world industrial environments, where the intended application of these methods occurs.

Intelligent fault diagnosis

Overview

Intelligent fault diagnosis (IFD) refers to the application of advanced techniques, often involving machine learning (ML) and deep learning (DL), for the purpose of identifying, analyzing, and diagnosing faults or abnormalities in a system. This can be applied across various domains, including industrial machinery, electrical systems, manufacturing processes, and more. IFD leverages intelligent algorithms to automatically analyze data, detect deviations from normal behavior, and identify the root causes of faults.

In the context of Industrial Machinery Health Management (IMHM), intelligent fault diagnosis plays a crucial role in ensuring the optimal functioning and reliability of industrial equipment. The key aspects of its role include early fault detection, condition monitoring, root cause analysis, predictive maintenance, and data-driven decision-making. The process of IFD is composed of three basic components that contribute to the effectiveness and efficiency of fault diagnosis systems. These processes of data processing, model creation, and training optimization are researched in depth over different domains. Data processing involves preparing and pre-processing raw data collected from sensors or other sources to make it suitable for analysis and modeling. Model creation involves designing and building intelligent models that can analyze processed data to detect faults and anomalies. Lastly, training optimization focuses on enhancing the efficiency and performance of the model during the training phase.

Data processing techniques

Feature extraction

Feature extraction and data fusion play crucial roles in intelligent fault diagnosis by enhancing the quality of information used by diagnostic models. Data fusion involves combining information from multiple sources or sensors to create a more comprehensive and accurate representation of the system under consideration. In fault diagnosis, different sensors or data sources may provide complementary information. Data fusion allows the integration of diverse data modalities, such as vibration sensors, temperature sensors, and acoustic sensors, to create a more holistic view of the system's condition. In scenarios where multiple sensors are distributed across a system, data fusion helps aggregate information from various locations to form a coherent understanding of the overall system's health. On the other hand, feature extraction involves transforming raw data into a set of relevant, informative features that capture the essential characteristics of the system. Fault diagnosis often deals with high-dimensional data. Feature extraction methods help reduce the dimensionality by selecting or

transforming features while retaining critical information. Effective feature extraction enhances the model's ability to recognize patterns associated with normal and faulty states. Extracting relevant features can help filter out irrelevant or noisy information, improving the signal-to-noise ratio in the diagnostic process. Huaxiang Pu et al. proposed a restricted sparse network (RSN) that can learn interpretable and lightweight features from vibration data. The network uses a novel frequency-domain space, a multichannel fusion mechanism, and a high-power feature extraction module to achieve high accuracy and reliability[68]. Wan et al. have proposed a multisensor information coupling network (MICN) that can process the data from different types of sensors, such as vibration, current, and torque, and fuse their features layer by layer using a mutual attention mechanism[69]. The authors have developed a feature information coupling algorithm based on a mutual attention mechanism, which can calculate the attention weights between the features of different sensors and reconstruct the features with the fused information. Meng et al. have proposed a lightweight model based on depthwise convolution and uses a multiscale features module for low-cost key feature extraction[70]. Ma et al. proposed a model, which consists of an efficient feature extractor (MiniNet), two domain-specific modules, and a classifier[72]. The model also uses a balanced strategy and a discriminative feature learning strategy to optimize the diagnosis performance. Buchaiah et al. have presented a framework for extracting significant features from bearing vibration data using various data fusion techniques and applying them to fault classification and remaining useful life prediction[74]. The authors extract seventy-two original features from the vibration data using multiple signal processing techniques and select a relevant subset of features using the Random Forest method to reduce complexity and overfitting. A two-step framework is proposed[59], first features are extracted and mapped using an adversarial autoencoder. Then the mapped features are clustered based on a Gaussian mixture model. Mao et al. has come up with an approach that seeks the optimal core tensor and domain-invariant features to later integrate with information processed by a pre-trained network to make accurate predictions[67]. Zhang et al. have proposed a novel fault diagnosis approach for multivariate data based on multivariate dynamic mode decomposition (MDMD)[54]. The approach converts the multivariate fault data into a tensor format and applies MDMD to extract the fault characteristic frequencies. The framework is based on the assumption that the fault features are low-rank distributed in the dynamic system tensor. Liu et al. have proposed a novel intelligent fault diagnosis method for rolling bearings under nonstationary conditions. The method uses a flexible generalized demodulation technique to transform the time-varying frequencies of vibration data into constant values that are related to the base frequency and the fault characteristics[57]. The method can overcome the effects of operation conditions on the demodulation results, and can automatically extract effective and robust features for fault recognition. Yan et al. have designed a deep learning model that can fuse the information of multiple shaft vibration images and learn fault features in an end-to-end manner[73]. Brusamarello et al. have presented an automatic system for detecting different fault types and severity levels in the outer bearing's raceway[56]. The authors have explained how the dynamic data measured by the sensors are pre-processed, transformed, and reduced using fast Fourier transform, power spectral density, and feature extraction methods. Yang et al. proposed a method of triplet network, which aims to learn the high-order features of vibration data and classify them into different fault categories. The method consists of two steps: feature extraction using a triplet network based on a one-dimensional CNN model, and fault identification using the classifier. The method uses a triplet loss function to minimize the distance between samples of the same class and maximize the distance between samples of different classes in the feature space[75].

Attention networks play a crucial role in data processing, particularly in tasks that involve handling and understanding complex information. The attention mechanism allows to focus on specific parts of the input data that are deemed more relevant for the task at hand. In sequential data processing, such as natural language processing or time-series analysis, attention networks help capture dependencies across different parts of the sequence. Meng et al. used high-speed cameras and acoustic sensors to collect vibration and sound data from bearing systems in different health states. They have proposed a method to fuse the heterogeneous data using structural modal analysis and graph neural networks[61]. The authors have used a graph attention network to learn the features and relationships of the fused data. The graph attention network can assign different weights to different nodes and edges based on the feature similarity and the structural information. A new convolutional neural network model that combines a convolutional attention mechanism and an improved ResNet18 network is proposed in Ref.[62]. It can extract useful features from vibration data and classify different types of bearing faults. The 1-D vibration data are converted into 2-D gray-scale images using MATLAB and then fed into the CBAM-ResNet model for feature extraction and fault diagnosis. The proposed method is tested on two public datasets (CWRU and XJTU) and compared with several other deep-learning models. The results show that the proposed method has higher accuracy, lower training time, and smaller model sizes than the other models. Xue et al. have proposed a method based on a self-calibrated coordinate attention mechanism and multi-scale convolutional neural network for rolling bearing fault diagnosis under small sample conditions[63]. This method converts vibration data into 2D images and extracts fault features automatically. The attention mechanism can capture both channel and location information of the features and enhance the classification ability of the model. The proposed deep learning method[64] combines self-calibrated convolution and split-attention mechanism to handle fault diagnosis and prediction using imbalanced data of rotating machinery. The authors have reported the prediction and fault identification accuracy of their method for the planetary gearbox dataset. They have shown that their model can achieve high accuracy under different imbalance ratios and demonstrated the feasibility of their method. Wang et al. have proposed a data-driven approach that combines features from different types of data to enhance fault diagnosis performance. Features can be extracted from vibration, acoustic, current, or other data types using methods such as CNN, wavelet transform, or STFT[65]. This approach is based on two novel techniques to achieve effective feature fusion. Mutual attention allows features from different modes to interact and guide each other, while the bilinear model performs fine-grained fusion of features from different sensors. The attention module proposed in Ref.[66] is used to enhance the discriminative capability of multiscale features by capturing the relations among different scales and channels. This module assigns different weights to different

scales based on their importance for classification. Lee et al. have proposed a novel model called MSCNet, which combines filtering methods, multi-scale feature extraction, and residual attention mechanisms to identify bearing faults under strong noise and varying speed conditions[71]. The paper introduced a morphological filter and a mean filter to preprocess the vibrational data and extract impulse and noise-reduced features. A multi-scale residual attention and multi-channel network (MSCNet) is designed, to learn from the filtered data and focus on the key features at different scales. The paper claims that MSCNet achieves better performance than five state-of-the-art networks on two-bearing datasets. Xiaotian Zhang et al. proposed an inferable deep distilled attention network (IDDAN) method, which uses a self-attention mechanism and a transfer learning technique to diagnose and classify multiple bearing faults in different motor drive systems[58]. Data augmentation is used on 1-D vibration data to provide large-scale pretraining samples for the network.

Data augmentation

These are special-purpose neural networks with configurable architecture to achieve multipurpose objectives. In the published research most commonly used adversarial network is GAN (generative), which focuses on generating synthetic data that is indistinguishable from real data. The adversarial training in GANs involves a generator creating data and a discriminator determining whether the data is real or generated. The researchers have used another variation known as Domain Adversarial network (DAN), whose primary objective is to learn a feature representation that is domain-invariant, meaning it captures the underlying patterns in the data that are consistent across different domains. Li et al. have introduced a novel domain adversarial network (DAN) guided by probability, capable of generating uncontaminated semantic features and achieving improved cross-domain diagnosis[77]. High deterministic domain adaptation is achieved through the integration of a classifier discrimination module. To handle the problem of misclassification authors have added a classifier discrimination module. The proposed method is tested on data generated by a test bed at Soochow University, improved diagnostic efficacy of 19.10%, 15.20%, and 13.7% is achieved in comparison to batch norm maximization, deep CORAL, and maximum classifier discrepancy respectively. Baokun Han et al. have proposed a domain adaptive fault diagnosis method based on the Wasserstein generative adversarial network (WGAN) and stacked autoencoders (SAEs) to deal with the problem of imbalanced data in bearing fault diagnosis. The authors use WGAN to generate balanced data from unbalanced data of different bearing fault types. They claim that WGAN can produce high-quality data that preserves the frequency domain features of the original data. They have used SAEs to extract fault features from the source and target domains and add L2 regularization to prevent overfitting[78]. The authors conduct two experiments to verify the effectiveness of their method under stable and variable speed conditions. They compare their method with five existing methods and show that their method achieves the highest accuracy, stability, and efficiency in all transfer tasks. Liu et al. has proposed a condition multidomain generative adversarial network that generates synthetic fault samples from limited real data and improves the accuracy of fault diagnosis[26]. This method introduces a self-adaptive feature extraction module to construct the sample condition information and a self-attention mechanism to capture the local and global fault features. Adversarial networks can be built in different ways, Ren et al. have used dual classifier discriminator adversarial networks that can separate new fault types from shared health types in the target domain by learning credible weights for the target samples. The model also uses a parallel channel attention module (PCAM) to extract key health state information from noisy vibration data[80]. The authors claim that their model has a novel weighting strategy that considers the similarity between the source and target domains from different perspectives, and a new PCAM that combines max pooling, average pooling, and original data in a parallel structure to enhance feature extraction capability. In real industrial scenarios, a common problem is facing a new unknown type of fault that may not have occurred before. Researchers have put a lot of focus on this problem and used different cross-domain learning techniques. Su et al. have proposed two methods, a target domain slanted classifier (TDSC) that leverages the data distribution of the target domain to overcome the biased learning problem, and an adaptive threshold that enhances the distinguishability between known and unknown faults based on the posterior probability of TDSC[81]. Shao et al. have proposed a dual-threshold attention-guided generative adversarial network (DTAGAN) to generate high-quality infrared thermal images that can assist in fault diagnosis. The method consists of three components: an improved loss function based on Wasserstein distance and gradient penalty, an attention-guided GAN to extract global thermal features, and a dual-threshold training mechanism to improve generation quality and efficiency[76]. Luo et al. proposed an intelligent method based on Wasserstein GAN Meta-Learning (WGANML), which can generate missing samples and enhance weak features of early bearing faults. It also uses meta-learning to train the network parameters without manual intervention[40]. Basic theory and method flow of the original GAN model and meta-learning model are introduced in the article with the bearing structure and fault features in different scenarios.

Zhenglin Dai et al. introduce Con-GAN, a novel improvement of GAN-based data augmentation that generates continuations of existing data and merges them with the original data[83]. Chen et al. propose a lightweight and robust model for engineering cross-domain fault diagnosis via feature fusion-based unsupervised adversarial learning[84]. The model can extract domain-invariant features from different working conditions and noise levels and has extremely small volumes and computation. Liu et al. present a novel fault diagnostic approach that incorporates an encoder into the discriminator to extract deep features of the original samples and uses a variational information constraint technique to improve the training stability and convergence of the generative adversarial network[82]. The paper also adds a representation matching module to avoid the mode collapse problem and boost the sample diversity. Mode collapse is a common issue in generative modeling, particularly in the context of training GANs. It refers to a situation where the generator of the GAN fails to produce a diverse set of samples and instead collapses to generate a limited set of similar or identical samples.

Imbalanced learning

Imbalanced learning methods are techniques specifically designed to address the challenges posed by imbalanced datasets, where one class is significantly underrepresented compared to others. In the context of intelligent fault diagnosis (IFD), imbalanced datasets can lead to biased models that perform poorly on minority classes. In recent years research community has specifically focused on methods to improve data balance for higher classification accuracy in data-driven systems. Li et al. in their research have used dual-stage attention-based recurrent neural network (DA-RNN) to extend imbalanced datasets[85]. Zhang et al. have proposed a novel scheme that uses transfer learning based on representation learning for few-shot fault classification, the method uses a self-attention-based autoencoder that uses the healthy data for training i.e. only a few fault samples for offline training and achieves high-accuracy online bearing fault classification[86]. Liu et al. have proposed a three-stage method, a domain-shared feature extractor based on a deep residual network, a cross-domain transfer learning module, and an explicit weight self-assignment module based on metadata[87]. The method aims to leverage the prior knowledge from the source domain, reduce the distribution discrepancy between domains, and rebalance the sample weights to improve the fault classification performance. Ding et al. have addressed the problem of imbalanced domain adaptation, which assumes that both feature shift and label shift exist between the source and target domains and that the label distribution is imbalanced in one or both domains[88]. This is the proposed framework that combines cost-sensitive learning, categorical alignment, and margin loss regularization to overcome the challenges of IDA and improve cross-domain fault diagnosis accuracy. Wang et al. have crafted spectrum distance to quantitatively measure the disparity in spectrum location between authentic data and generated data. This distance metric serves as a guide during model training, facilitating the generation of high-quality data that closely resembles the features present in authentic data[89]. Balancing data in IFD is crucial for training models that can effectively handle imbalanced datasets. However, there are challenges and limitations associated with it like overfitting to minority class, loss of information, increased model complexity, and sensitivity to sampling techniques. To ensure application of most suitable strategy in the scenario of imbalanced learning scenarios it important to distinguish between outliers and rare samples, thereby improving model robustness and accuracy. Outliers are data points that deviate significantly from the rest of the dataset, often representing noise or erroneous entries. They can negatively impact the performance of machine learning models by introducing bias or misleading the training process. Common techniques to handle outliers include outlier detection methods like isolation forests, z-score analysis, or robust statistical techniques, which aim to identify and potentially remove these data points to prevent model distortion. On the other hand, rare samples represent legitimate but infrequent classes within the data, often associated with minority class instances in imbalanced datasets. Unlike outliers, rare samples are crucial to the learning process and should not be removed or ignored. To handle rare samples, oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) or undersampling of the majority class are commonly employed to balance the dataset. Additionally, cost-sensitive learning and ensemble methods like boosting can improve the model's sensitivity to rare classes by assigning higher importance to these instances during training.

Noise removal methods

The noise in vibrational data can affect the processing capability of diagnostic algorithms. Noise removal in time-series data is a crucial step to enhance the signal-to-noise ratio, making underlying patterns more discernible. Researchers have used various techniques for noise removal in time-series data. The utilization of the recurrence plot (RP) method in bearing fault diagnosis is attributed to its effectiveness in analyzing nonlinear and non-stationary data within dynamic systems. However, noise inference can substantially reduce the effectiveness of this method. Liu et al. have proposed an anti-noise recurrent plot-based method that is supported by CNN which can deliver accurate fault diagnosis[90]. Their results exhibit a 90% accuracy under Gaussian white noise with a 6+ dB signal-to-noise (SNR) ratio. Vibration data is recursive in nature which helps in constructing recurrence graphs (RGs). Yuan et al. have proposed an algorithm that transforms vibration data from time series to graph representation. The authors have introduced spatial noise to overcome the overfitting of the resulting multichannel graph convolutional network (GCN)[91]. Wang et al. has proposed a multi-scale attention network with adaptive noise reduction to deal with the complex and noisy data of aero engine bearings. The model consists of four modules: multiscale partitioning, multiscale noise reduction, attention fusion, and classifier[66]. The core of the noise reduction module is the threshold noise reduction (TNR), which uses global average pooling (GAP) and max pooling (MAP) to extract the periodic and impulsive characteristics of the fault data, and then learns an adaptive threshold to suppress the noises. Lyu et al. have proposed a method that combines the functionalities of a soft threshold and global context to accomplish efficient noise reduction and feature extraction[92]. Soft thresholding is effective in removing noise while preserving important features by setting values to zero which fall under the threshold. Noisy labels and the environment, are the two main challenges that the proposed method[93] aims to overcome. Noisy labels refer to the mismatch between the true class and the training label of a sample, and noisy environment refers to the interference of signal noise in the acquired data. The proposed method uses a new loss function with two adjustable parameters to deal with noisy labels and shows better robustness against noise than other methods. Chengjiang Zhou et al. has innovated an improved VMD based on mutual information to suppress noise[24].

Multi-scale time-series analysis

The five generic steps of a fault diagnosis method involve data collection, decomposition, feature extraction, feature fusion, and fault identification. This technique involves calculating the spectrum of the envelope of a signal. In fault diagnosis, particularly in machinery, the envelope spectrum can highlight specific frequency components associated with faults, making it easier to identify and analyze issues. A portion of research has been dedicated to this domain of techniques for fault diagnosis and prediction. Hua et al. have proposed a

method for bearing fault diagnosis based on modified frequency band envelope kurtosis (MFBEK), which uses wavelet packet transform, envelope kurtosis, correlation coefficient, and band-pass filter to enhance the fault feature extraction and noise reduction performance[96]. Li et al. have proposed a fast and adaptive empirical mode decomposition (FAEMD) method, which can decompose nonstationary and nonlinear signals into intrinsic mode functions (IMFs) using order statistics filter (OSF) and adaptive window width[97]. Chen et al. have proposed a new feature extraction method based on hierarchical improved envelope spectrum entropy (HIESE), which can capture the complexity and frequency information of bearing vibration signals[98]. The authors have used support vector machines (SVMs) as a fault identification model. A new blind deconvolution method[99] for rotating machinery condition monitoring is proposed. The method is based on maximizing the spectral harmonics-to-interference ratio (SESHIR) of the filtered signals, which can better detect and extract the fault impulses caused by mechanical defects. Renxiang Chen et al. have also used blind deconvolution with cyclostationarity indexes to diagnose bearing faults[100]. Li et al. have addressed the drawbacks of the existing cyclostationary blind deconvolution (CYCBD) method, which requires manual setting of the cyclic frequency and filter length. They develop a novel method that can estimate these parameters adaptively and automatically by using noise subtraction, autocorrelation envelope, and residual autocorrelation energy ratio[101]. Yi et al. have proposed an adaptive harmonic product spectrum method for rotating machinery fault diagnosis, which can locate the optimal resonance frequency band based on signal power spectral density and harmonic saliency index[102]. Pan et al. introduce a new matrix classifier based on the probability framework and symplectic geometry theory for roller bearing fault diagnosis[103]. Ma et al. propose a novel method called optimized periodic mode decomposition (OPMD) to improve the extraction of periodic components from acquired data[104]. The OPMD method uses phase compensation and correlation detection to enhance the adaptability and accuracy of the Ramanujan subspace theory, which can decompose any signal into a series of periodic components.

Model creation with deep learning

Overview

Deep learning is a subfield of machine learning that involves artificial neural networks with multiple layers (deep neural networks). These networks are capable of automatically learning hierarchical representations of data by processing it through multiple layers of interconnected nodes. The depth of these networks allows them to capture intricate patterns and features, making them particularly effective in tasks such as sequential data processing, image recognition, speech processing, natural language understanding, and more. Deep learning has shown significant promise and relevance in the field of intelligent fault diagnosis which is reflected through a positive trend in research over recent years. Intelligent fault diagnosis involves identifying and diagnosing faults or anomalies in systems, machinery, or processes. Deep learning models excel at automatically learning relevant features from raw data. This is crucial in fault diagnosis, where the identification of subtle patterns or anomalies may be challenging for traditional methods. Faults in systems often exhibit non-linear and complex relationships. Deep learning models, with their ability to model complex functions through layers of non-linear transformations, are well-suited for capturing intricate patterns associated with faults that might be difficult to represent using traditional linear models. These models can perform end-to-end learning, meaning they can learn directly from raw input data to the final output (diagnosis) without the need for manual feature engineering. The types of faults and operating conditions vary in real industrial environments, leading to diverse data for which training a model is difficult. The adaptability and generalization capability of deep learning models makes them an ideal candidate to handle variation and scarcity of data. The data during industrial processes is collected from multiple sources and deep learning architectures can handle multimodal data integration making them a better choice in carrying IFD.

Neural networks

Neural networks mimic the human brain by interconnecting layered nodes into a decision-making network. Like the human brain neural networks learn with training by adjusting weights associated with neurons (connection between nodes). Neural networks are capable of learning complex patterns and representations from data, making them effective in various domains, including fault diagnosis. These networks can be trained to recognize normal system behavior, enabling them to detect anomalies or deviations indicative of faults. This is particularly useful for identifying incipient faults that may not follow well-defined patterns. In recent years, exponential growth in processing power and reduction in the size of microprocessors have driven the research community to focus on the use of neural networks in industrial machinery health management. In this study we have focused on feed-forward fully connected neural networks as they have the ability to model complex relationships between input features and the output, i.e., detection and prediction. Liang et al. have proposed a multibranch and multiscale dynamic convolutional network (MBSDCN) that can extract features from small samples of vibration signals and classify different fault types of rolling bearings and gearboxes. The authors have introduced the multi-branch and multiscale dynamic convolutional layer (MBSDCL) module, which splits the input into multiple branches, applies multiscale convolution to each branch, and calibrates the weights of each convolutional layer by a channel reconstruction (CR) attention mechanism[111]. Liu et al. have proposed a multiscale fusion attention convolutional neural network (MSFACNN) method, which consists of two parts: data preprocessing and fault diagnosis[112]. The data preprocessing converts the 1-D vibration signals into 2-D grayscale images, which preserves the time-domain information. The fault diagnosis uses a MSFACNN model, which has a multiscale feature extraction module, and a spatial attention module. The MSFACNN model can automatically learn the abstract representation of the input data and focus on the key fault features. Chang et al. have proposed a semi-self-supervised method (S3M) that consists of two learning stages: self-supervised learning and supervised learning. S3M can use both labeled and unlabeled data to learn robust and discriminative features for fault diagnosis[113]. The authors use convolutional neural networks as the

basic structure for the feature extractors and the classifiers in S3M and they have designed two auxiliary tasks based on contrastive learning and noise disturbance to extract global and local features of the vibrational data and to encourage feature separability between different classes of samples. The pretext tasks are implemented by three feature extractors and two classifiers. Niu et al. have proposed a deep learning-based method for bearing fault diagnosis and localization with information fusion and feature attention[55]. The paper explores the attention mechanism in multitask deep residual (DR-CNN) to improve bearing diagnosis performance. Zhao et al. have constructed a novel convolutional deep belief network with Gaussian distribution, which can overcome the variance of the conventional convolutional deep belief network and improve the feature learning and fault classification performance[106]. Xue et al. have proposed a local binary temporal convolutional neural network (LBTCNN) for bearing fault diagnosis[107]. The network consists of a local binary convolution (LBC) layer, a series of 2-D convolutional layers, a residual flow, a temporal module, and a softmax layer. The LBC layer preserves the local features, the temporal module captures the temporal dependencies, and the residual flow avoids feature loss. Zhang et al. propose a method that combines data probability density, gram angular field, and convolutional neural network (DPD-GAF-CNN) to convert one-dimensional vibrational data into two-dimensional images and classify them using deep learning[109]. The authors claim that their method can simplify the feature extraction process, reduce the dependence on expert knowledge, improve the accuracy and stability of fault diagnosis, and adapt to different signal lengths and working conditions. Li et al. have proposed a lightweight neural network that uses fewer computational resources and achieves higher accuracy than existing methods. It combines the feature extraction capability of CNN and the fitting ability of multi-layer perceptron (MLP) with 1×1 convolution kernels[110]. The research involves a loss function that adjusts the data distribution and limits the weight range during training to improve the robustness and generalization of the model. Metrics used for comparison are accuracy, training time, and model complexity. A deep belief network (DBN) is a type of artificial neural network that is composed of multiple layers of stochastic, latent variables. It is a generative probabilistic model that consists of two main components: the Restricted Boltzmann Machines (RBMs) and a top-level layer known as the visible layer. DBNs are often used for unsupervised learning tasks, such as feature learning, dimensionality reduction, and generative modeling. Haihong Tang et al. propose a novel method for bearing fault diagnosis using a minimum unscented Kalman filter-aided deep belief network (MiUKF-aided DBN) that can extract invariant features from vibrational data collected by multiple sensors[114]. The feature maps can effectively transform the time-series data from multi-sensors into 2-D feature maps that preserve the temporal relation and the comprehensive fault information.

Convolution neural networks

Multiscale graph convolution neural networks (MG-CNN) can improve the effectiveness and accuracy of fault detection and prediction. These neural networks can capture multilevel dependencies, enable hierarchical feature extraction, bring robustness to handle scale variations and improve generalization. Peizhe Yin et al. have proposed a multiscale framework based on MG-CNN to extract features from vibration data at different scales and adapt them to the same dimension[123]. They proposed to integrate the features obtained at single scales with the features obtained at a cross-scale using a multiscale graph iteration module. A mutual fusion method based on the Bayesian model is also presented in the research to fuse the features from different scales and achieve better fault classification accuracy. Feng et al. have designed a domain adversarial graph network (DAGN) that uses graph convolutional networks to extract and align features from different domains and achieve effective fault diagnosis of bearings[119]. The graph network is digital twin-enabled and can transfer knowledge from simulated data to measured data to detect bearing faults with limited knowledge. A digital twin refers to a virtual representation or simulation of a physical object, system, or process. This concept involves creating a digital counterpart that mirrors the real-world entity in a virtual environment. The digital twin can replicate the physical entity's characteristics, behavior, and interactions, providing a comprehensive and dynamic representation. Mitra et al. have proposed a fusion method of adaptive superlet transform (ASLT) and 2D-CNN to diagnose the health condition of bearings in induction motors using vibration data. ASLT is a time-frequency transformation that provides high resolution and super-resolution for non-stationary data. 2D-CNN is a deep learning architecture that can extract relevant features from images and classify them into different categories[105]. The authors use a laboratory-scale industrial process model with two induction motors, one of which is the primary actuator whose bearing health is monitored. The vibration data is collected from the primary actuator under different speeds and external vibrations using a piezoelectric sensor and a data acquisition system. The acquired data are segmented and transformed into RGB images. Li et al. proposed a novel method that combines image processing techniques, graph structure construction, and semi-supervised graph convolutional network to achieve high-accuracy fault diagnosis of rotating machinery with few labeled samples[122]. The article uses the normalized cross-correlation coefficient (NCC) to measure the similarity of the images generated by each sample and determines the adjacency relationship between the nodes based on the similarity threshold. The graph structure reflects the connection and difference among the samples. Li et al. a novel deep learning architecture, named Adaptive Multiscale Fully Convolutional Network (AMFCN), for intelligent bearing fault diagnosis under various noise environments[116]. The AMFCN uses random sampling, huge kernel convolution, and adaptive multi-scale convolution to enhance the feature extraction ability, noise immunity, and robustness of the network. Zhao et al. proposed a new algorithm called multiscale deep graph convolutional networks (MS-DGCNs), which combines a new intra-class fine-to-coarse multiscale fusion technology and multiscale graph convolution kernels[118]. The article presented an intelligent fault diagnosis method based on MS-DGCNs for the rotor-bearing system under fluctuating conditions. Yang et al. has proposed a novel called balanced deep transfer network (BDTN), which uses a modified CNN as the backbone and aligns both the marginal and conditional distributions between the source and target domains using maximum mean discrepancy (MMD) and pseudolabels. A balancing factor is introduced to adjust the relative importance of the two distributions

dynamically[121]. Cui et al. have proposed a method that uses data fusion and CNN to detect defaults from vibrational data collected from multiple bearings[108]. The data is first decomposed into components such as amplitude, phase, and eccentricity. Features are derived from these components using the instantaneous orbit technique. CNN is used to learn and then classify fault types. Transfer learning is added in the end to reduce training time and improve accuracy. Lu et al. proposed a novel fault diagnosis method based on spectrum alignment (SA) and deep transfer convolution neural network (DTCNN) for unbalanced bearing data under various speeds[120]. SA algorithm transforms nonstationary vibrational data into stationary data by eliminating the influence of speed fluctuation and aligning the spectrum of all samples based on the first sample. A data augmentation module that uses DBSCAN clustering and an unbalanced sampler to balance the data and reduce the overfitting of the model. Finally, a deep learning model that uses two loss functions, cross-entropy and transfer regular term, learns domain-invariant features, and achieves fault classification. Chunran Huo et al. propose a one-dimensional convolutional neural network with a stronger linear fitting ability with a feature-based transfer learning method that uses pseudo-labels of the target domain to guide the network to learn the data features of the target domain[117].

Recurrent neural networks

Recurrent neural networks (RNNs) are a type of neural network architecture designed to work with sequential data by maintaining a hidden state that captures information about previous elements in the sequence. Unlike feedforward neural networks, RNNs have connections that form directed cycles, allowing them to maintain and update a hidden state as they process each element in the sequence. In the context of IFD, RNNs play a crucial role in handling sequential data related to the operation and behavior of machinery. Its main contributions are time-series analysis, anomaly detection, predictive maintenance, sequence-to-sequence modeling, and handling temporal dependencies. The exponential growth of computing power has enabled the use of RNNs in real-time scenarios and complex situations which has led to hyped research interest in this domain. Zhu et al. have conducted a literature review on the application of RNN to machine fault diagnosis before 2022[125]. Imamura et al. have proposed a low-delay lightweight recurrent neural network (LLRNN) that uses either LSTM or JANET cells to identify the unbalance of light rotating machinery, using mechanical vibration or motor current signals[126]. The authors have shown that it is possible to reduce the computational cost of the RNN by using fewer memory blocks with one cell each while maintaining high accuracy and performance. They also demonstrated the feasibility of using electric current data for unbalance diagnosis, which gives more flexibility for industrial applications. Chang et al. proposed a novel method for data-driven predictive analytics of rotating machinery, based on a fusion health indicator (Fusion-HI) and a heterogeneous bi-directional gated recurrent unit (GRU) model[127]. The Fusion-HI integrates multi-domain features to reflect the degradation state of equipment, while the heterogeneous bi-directional GRU model captures the complex non-linear mapping relationships of the time sequences. Zhiqiang et al. presented a novel method for fault diagnosis of rolling bearings under multiple working conditions using a deep neural network (DNN) with an attention gate and a multiscale recursive fusion strategy. The method aims to overcome the information loss problem in DNN and to extract the potential features related to working condition changes[128].

Auto encoders

Autoencoders are a type of artificial neural network used in unsupervised learning and dimensionality reduction. They consist of an encoder and a decoder, which are trained together to learn a compact representation of the input data. Key components of autoencoders are the encoder, decoder, and loss function. Autoencoders are used in data compression, anomaly detection, feature learning, and denoising. Autoencoders can play a significant role in the analysis of vibrational data, particularly in tasks related to feature learning, data compression, and anomaly detection. Sun et al. propose a new deep regularized autoencoder model that combines global and local neighborhood graphs with sparse graph embedding to extract fault information from data[141]. The model aims to improve the performance and generalization of the original autoencoder algorithm by introducing manifold and sparse regularization terms. Shi et al. have proposed a deep hypergraph autoencoder embedding (DHAEE) algorithm and a fault diagnosis approach based on it for rotating machinery with unlabeled data[142]. The method converts vibrational data into hypergraphs, uses hypergraph convolutional and self-representation layers to learn higher-order and subspace structural information, and stacks multiple hypergraph convolutional ELM autoencoders to extract deep abstract features. The graph embedding-based deep broad learning system (GEBLSAE)[143], rooted in the autoencoder framework and supervised graph embedding theory, is designed to enhance the extraction of efficient feature representations from data. Leveraging the autoencoder mechanism, it captures intricate features, while the supervised graph embedding framework strengthens the discriminative power of these representations. This approach involves learning the distinct flow structures associated with different classes, contributing to more effective and nuanced feature extraction. A few studies have focused on specific faults in rolling element bearing caused by specific conditions. In the event of a localized fault in the rolling element bearing, alterations in contact stiffness between the rolling element and the raceway at the fault position can result in consistent fluctuations in the instantaneous angular displacement (IAD) of the rotating shaft. This characteristic imbues the encoder with the capability for diagnosing faults in the bearings[144].

Long short-term memory (LSTM) methods

Long short-term memory (LSTM) networks are a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing long-range dependencies and handling vanishing or exploding gradient problems. LSTMs have gating mechanisms that allow them to selectively remember or forget information over extended sequences, making them particularly effective for tasks involving sequential data. The research community is focusing on LSTMs due to their inbuilt capability of modeling and analyzing

time-series data, for predicting faults in industrial machines and rotating components. Xu et al. have presented a novel method for detecting faults in wind turbine bearings using a multi-scale convolutional neural network with bidirectional long short-term memory (MSCNN-BiLSTM) and a weighted majority voting rule for multi-sensors[133]. The MSCNN-BiLSTM model uses an improved multi-scale coarse-grained procedure algorithm to capture multi-scale time information from raw vibration data of the bearings. Shi et al. have proposed a deep neural network based on bidirectional-convolutional long short-term memory (BiConvLSTM) that can extract spatial and temporal features from multiple sensor sources automatically and simultaneously, without losing critical information[134]. Yiyao et al. used a long short-term memory (LSTM) network to capture the long-term dependence features in the fault signals, and improve the accuracy of the diagnosis method[135]. Tang et al. used entropy gain ratio (EGR) and semi-supervised transferable LSTM network for fault diagnosis under variable working conditions[136]. The LSTM transfers the knowledge learned from the source domain to the target domain and fine-tunes the parameters of the output layer with a small number of labeled samples in the target domain.

Transformer-based methods

Transformers can be applied to capture patterns and dependencies in sequential data for time series forecasting. Zhu et al. propose a Periodic Representation for Transformers (PRT) model, which can extract features from raw vibration data of rotating machines and classify different faults[137]. The model uses dense overlapping patch splitting, class attention, and a two-stage positional encoding method to enhance feature learning. The authors train the model on small-size samples and test it on large-size samples, which can improve the performance and make more efficient use of the limited data. In addition, researchers have proposed methods reliant on vision transformers based on the fact that computer vision tasks are not dependent on convolution structure, and attention-based transformer models may directly be applied. Xu et al. have proposed a one-dimensional vision transformer encoder method for fault diagnosis[138]. This model takes one-dimensional data as input without any spatio-temporal domain conversion. Fang et al. have proposed a novel framework called X-self-attention convolution neural network (XACNN), which combines the global feature extraction ability of Transformer and the inductive bias ability of CNNs[139]. The research introduces a preprocessing method based on absolute value fast Fourier transform (AVFFT) to improve the data quality and the network accuracy. The authors have demonstrated the effectiveness and superiority of the proposed method on a self-made bearing fault diagnosis dataset and deployed the trained model on a smartphone as a portable fault diagnosis device. The authors have claimed that this is the first attempt to build a mobile intelligent fault detection device based on deep learning and provides a practical and low-cost inspection scheme for the industrial field. Haiyue Wu et al. introduce the Transformer model and its essential elements, such as the attention mechanism, the encoder-decoder structure, and the positional encoding[140]. The paper explains how the Transformer model can efficiently extract features and learn patterns from the spectrograms.

Ensemble learning

Gwak et al. has proposed a framework for diagnosing faults in industrial facilities using vibration data and deep learning models. The framework consists of two components: a power-perturbation-based decision boundary analysis (POBA) and a robustness-based ensemble (ROE) algorithm[145]. The ROE algorithm combines the prediction scores of multiple trained models to produce a robust prediction. The ROE algorithm uses the decision boundary information obtained by the POBA to evaluate the robustness score per class for each model. A large weight is then assigned to the prediction score for the class with a large robustness score. The ROE algorithm can achieve higher test accuracy than a basic ensemble model when the operating condition of the test data differs from that of the training data. Chen et al. have proposed a new predictive maintenance strategy that combines system failure prediction with maintenance and inventory decisions. The strategy uses a novel ensemble model of DAE, LSTM, QR, and KDE to obtain the probability density of system failure time, and then optimizes the replacement cost function and the ordering cost function to determine the optimal maintenance and inventory plans[146]. DAE as a deep neural network extracts low-dimensional features from raw data, while LSTM as a recurrent neural network learns the temporal correlation information in time series. Hongwei et al. have proposed a composite fault diagnosis model combining a convolutional neural network and a support vector machine (SVM)[147]. The model replaces softmax with SVM as the CNN classifier and compared to softmax, SVM employs a maximum margin classifier to differentiate sample data, aiming for theoretically higher diagnostic accuracy. The successful deployment of Support Vector Machines (SVMs) in fault diagnosis encounters various implementation challenges.

Advanced methods

In recent years, advancements in intelligent fault diagnosis (IFD) have increasingly focused on leveraging nonlinear methods and optimizing data-driven approaches for more robust and scalable solutions. These techniques, such as kernel-PCA and autoencoder-based fault detection, have demonstrated significant promise in enhancing accuracy and computational efficiency in various industrial applications. For instance, Learnable faster kernel-PCA for nonlinear fault detection[219] introduces a more efficient approach to nonlinear dimensionality reduction, while Optimized design of parity relation-based residual generators focuses on enhancing fault detection by optimizing residual generation techniques. The authors have proposed a learnable, faster realization of KPCA through a depp autoencoder-based feature extraction framework, termed DAE-PCA. The method automatically optimizes the selection of nonlinear high-dimensional spaces, enhancing accuracy in fault detection. Traditional fault diagnosis systems often overlook the optimal selection of parity vectors, which can limit their effectiveness. Yuchen et al.[220] introduces novel approaches to derive parity vectors that optimize residual generators for both linear and nonlinear systems. By analyzing the parity space dimension, a parameterization of all parity relation-based residual generators is proposed. An iterative process is used to find

the optimal parameters, minimizing regression error. Unlike conventional methods, which are primarily suited for linear systems, the approach is extended to handle strong nonlinearities using data-driven Hammerstein function estimation. The optimized residual generation algorithms are designed for both offline and online use. Evaluation on systems such as a three-tank system and a hot rolling mill process demonstrates that the proposed methods significantly improve fault detection sensitivity, especially for small faults, compared to traditional non-optimized approaches. Deep Responsible Active Learning (DRAL) refers to an advanced machine learning approach designed to address both data efficiency and model transparency issues in active learning scenarios. Active learning focuses on selecting the most informative data points for labeling, allowing models to learn effectively with fewer labeled examples. This methodology is highly applicable to industrial use cases, such as fault detection, where accurate and reliable decision-making is critical. Mahesh et al.[221] proposes an optimized one-dimensional Convolutional Neural Network (1D CNN) using multiple kernel sizes. This model is designed to automatically extract relevant features from raw sensor data, capturing complex patterns without manual feature engineering. The method outperforms existing approaches, achieving an accuracy of 99.52% in identifying bearing faults. The results highlight the proposed method's superior fault classification accuracy and its efficiency in detecting intricate fault patterns, making it a promising solution for real-time fault detection and predictive maintenance in bearings. To conclude, the incorporation of these advanced methods, demonstrates significant improvements in the accuracy and efficiency of fault detection systems. These approaches push the boundaries of current methodologies, providing deeper insights into nonlinear systems and complex data patterns, while enhancing real-time processing capabilities. Together, they represent key advancements in intelligent fault diagnosis, furthering the applicability of machine learning in industrial machinery monitoring.

Discussion

The application of deep learning (DL) models to industrial machinery has been motivated by the need to process complex, high-dimensional data effectively and deliver real-time insights for predictive maintenance and fault detection. Each DL architecture reviewed in the preceding sections brings distinct advantages and challenges, and understanding their suitability for specific tasks is essential for developing robust solutions. Neural networks (NNs), particularly feed-forward architectures, have laid the foundation for more advanced models due to their ability to learn abstract representations from data. However, their limited temporal awareness necessitated the evolution of more specialized architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs excel in feature extraction from structured data, particularly when working with vibration signals transformed into images (e.g., spectrograms). This makes them effective in fault detection tasks, but their lack of memory capabilities constrains their use for sequential data. RNNs, including their more sophisticated variant LSTMs, offer solutions to these limitations by capturing temporal dependencies in time-series data. This ability is crucial when analyzing degradation patterns over time, as it allows for more precise remaining useful life (RUL) prediction. However, they can be computationally expensive and prone to vanishing gradient problems with long sequences, which led to the emergence of transformer-based methods. These methods overcome sequence length limitations through self-attention mechanisms, enabling them to capture both short and long-range dependencies more efficiently. Autoencoders, often used for anomaly detection and dimensionality reduction, provide an unsupervised approach that is particularly valuable in industrial settings where labeled data is scarce. Ensemble learning, combining the strengths of multiple models, has proven to enhance performance, particularly when dealing with noisy or imbalanced data.

In this discussion, we emphasize that the choice of DL model depends heavily on the characteristics of the problem at hand. For industrial machinery, where fault detection, RUL prediction, and anomaly detection are key tasks, it is often necessary to strike a balance between model complexity and interpretability. While transformer-based models may provide the best results in capturing complex patterns, CNNs or LSTMs may be more efficient for real-time applications. This literature review highlights that no single DL architecture is a panacea for all predictive maintenance tasks. Instead, a hybrid approach-leveraging the feature extraction capabilities of CNNs, the temporal sensitivity of LSTMs, and the global attention mechanisms of transformers-offers a holistic solution to industrial fault detection and diagnosis. Table 5 highlights key parameters such as the number of layers, kernel sizes, learning rates, batch sizes, number of parameters, and configuration functions along with model-specific accuracy metrics. This comparison emphasizes the trade-offs between model complexity and performance, with models like MCBLS and Lifelong Learning Method (LLMGFR) demonstrating high accuracy with relatively efficient operation times. Through this analysis, we aim to offer a clearer understanding of the scalability and practicality of these models, thereby enhancing the value of this review paper by offering a more comprehensive evaluation of the model's performance.

Training optimization methods

Overview

Training optimization methods refer to techniques and algorithms used to improve the training process during intelligent fault diagnosis. The goal is to find the optimal set of parameters (weights and biases) that minimizes the error or loss function on a given dataset. Efficient training optimization is crucial for achieving better model performance, faster convergence, and improved generalization. The research literature has explored different optimization methods like stochastic gradient descent (SGD), which is a fundamental optimization algorithm used in training neural networks. It updates the model parameters by taking small steps in the direction of the negative gradient of the loss function with respect to the parameters. Momentum is an enhancement to SGD that introduces a moving average of past gradients to accelerate convergence and reduce oscillations. There are adaptive learning rate methods that dynamically adjust the learning rate during training based on the historical behavior of the gradients. Batch normalization is another method that normalizes the input to a layer during training, reducing internal covariate shifts and accelerating convergence. Dropout, randomly drops a fraction

Model/reference	Network parameters	Layers	Kernel sizes	Accuracy (%)
Domain adaptation network[162]	LR: 0.0001 BS: 128 E: 200	4,1	6 × 3 × 2 16 × 3 × 3 16 × 2 × 2 32 × 3 × 3 32 × 2 × 2	98.5
Lifelong learning method based on generative feature reply (LLMGFR)[163]	LR: 0.001, 0.0001 E: 200/400 BS: 128	4,1	64 × 3 × 3 128 × 3 × 3 256 × 3 × 3 512 × 3 × 3 512 × 4 × 4	99.03
1D-CNN(KD-MSFDA)[164]	Optimization: SGD LR: 0.01 (Automobile Transmission, AT) 0.05 (Rolling Element, RE)	4	128,5,5,15	AT: 95.64 RE: 90.18
2-D Convolution, MK-MMD[34]	LR: Pre-training: 0.00005 Training: 0.00003 BS: 128	6	4, 64, 128,256, 512, 1024	99.45
Mutual attention[65]	Parameters: 98499	4	–	99.86
CNN[110]	LR:0.001 Decay step: 3000 (bearing DS) 4000 (gear DS) BS: 50	3	4, 16, 64	99.44
Multiscale convolutional neural network[28]	LR: 0.001 BS: 128	7	16, 64, 16 8, 8, 16 16, 16, 16 32, 32, 16 64, 64, 16 32, 3, 16 64, 3, 16	DS 1: 99.10 DS 2: 96.47
Augmentation techniques using CNN[25]	Filter size: 3 × 3 LR: 0.001 BS: 32	4	64, 32, 16, 8	99.98
Attention-based RNN and self-calibration CNN[64]	Time window: 10 Layer size: 128, Hidden layer size: 128 LR: 0.001 BS: 32	4	3 × 3 7 × 7	99.25
Multi-sensor information coupling network (MICN)[69]	LR: 0.001 BS: 64	13	34 64 × 3 144 160 × 2 256 × 3 384 × 2	99.23
Self-supervised deep tensor domain-adversarial regression adaptation approach[67]	Optimizer: Adam LR: 0.001 WD: 0.00005 DR: 0.2 Time step: 8	13	[2560, 1024, 512, 128, 50] [50, 100, 50, 25, 1] [25, 100, 2]	–
Multiscale attention network (MANANR)[66]	Dropout: 0.5 Optimizer: Adam LR: 0.001 Loss function: cross-entropy	4	32,64 1066 × 64 512	99
De-formable space-frequency attention network (DSFAN)[23]	BS: 16 E: 50 LR: 0.003	4, 1, 3	100, 800 1, 3 × 3, 1, 1 8, 3 × 3, 1, 1 1, 3 × 3, 1, 1 8, 3 × 3, 1, 1	100
Instance weighting-based partial domain adaptation[165]	LR: 0.001 Epochs: 25, 40 BS: 64 Activation function: Softmax	4	[16, 1, 16] [8, 1, 32] [8, 1, 32] [100] Stride: [2,1]	Use case I: 89.65 Use case II: 96.28
Multibranch and multiscale dynamic convolutional network (MBSDCN)[111]	Epoch: 160 LR: 0.001 BS: 32	4	64,32,16,8 Channel: 16, 64 Stride: 8, 1	Bearing DS: 92.97 Gearbox DS: 96.94
Target domain slanted adversarial network (TDSAN)[81]	E: 1000 LR: 0.001 BS: 100	–	–	Bearing DS: 97.76 Gearbox DS: 95.82
Manifold-contrastive board learning system (MCBLS)[131]	–	Flat neural network	–	99.64
Multiscale fusion attention CNN (MSFACNN)[112]	Optimizer: Adam LR: 0.001 BS: 32	6	16 [3/2, 5/2, 7/2], 1/1 1 [3,3,3], 1	98.48

Table 5. Comparison of deep learning models in terms of network structure and accuracy.

of neurons during training, preventing overfitting and improving generalization. The benefits are that it acts as a form of regularization and reduces co-dependency among neurons. Schedulers are used to systematically decrease the learning rate over time during training which can help the model converge faster in the initial stages and fine-tune more precisely in later stages. Initialization is a key factor that helps prevent vanishing or exploding gradients during training, there are different strategies used to properly adjust the weights. Gradient clipping is another method that can stabilize training, particularly in recurrent neural networks and deep networks by limiting the magnitude of gradients during training. L1 and L2 regularization add penalty terms to the loss function based on the magnitudes of model parameters, preventing overfitting. Researchers have used meta-learning, transfer learning and ensemble methods also for optimization.

Hyperparameter tuning

Hyperparameter tuning is the process of systematically searching for the optimal set of hyperparameters for a machine-learning model. Hyperparameters are configuration settings that are external to the model itself and are not learned from the data during training. They control the learning process and influence the performance of the model. Hyperparameter tuning is commonly used in deep learning to optimize the performance of neural network models. Deep learning models, particularly neural networks, have numerous hyperparameters that influence their architecture, learning process, and generalization capabilities. Researchers have focused on techniques to optimize several different hyperparameters like learning rate, batch size, number of hidden layers and units, activation functions, weight initialization and regularization techniques. Lee et al. have presented extensive benchmarking results on the tuning of optimizer hyperparameters for convolutional neural network (CNN) models applied to machine fault diagnosis using raw vibration signals[148]. The paper sets the hyperparameter search space based on previous studies and then trains the models using hyperparameters sampled from a quasi-random distribution. The authors have evaluated model performances using noise-free and noisy data and compared the accuracy and computational efficiency of different CNN models, optimizers, and batch sizes. The key findings of the research are, the learning rate and momentum factor, which determine training speed, substantially affect the model's accuracy. The paper also discovers that the impacts of batch size and model training speed on model performance are highly correlated. Xingchen et al. have addressed the challenge of accurate bearing fault classification using machine learning, which depends on efficient features and optimal hyper-parameters[149]. The authors have proposed a new hyper-parameter optimization method for high dimensions based on dimension reduction and partial dependencies. The method can automatically construct domain features from raw bearing data and tune the hyper-parameters of both feature engineering and machine learning algorithms. Zhang et al. have proposed a hyper-parameter search algorithm, which introduces two convolutional neural networks (LeNet5 and AlexNet) that are used to find the optimal hyper-parameters and verify their effectiveness[150]. Hyperparameter tuning, while crucial for optimizing deep learning models, presents several research challenges that researchers continually strive to address. Some key challenges in the field of hyperparameter tuning research include, exploring and navigating high-dimensional search space poses a significant challenge, developing methods that balance computational cost with the need for thorough exploration is challenge and developing methods to transfer knowledge gained from hyperparameter tuning in one context to another, these are most recent research areas. Optimizing the parameters of a classifier using a metaheuristic algorithm is a common approach that falls under the category of metaheuristic optimization for hyperparameter tuning. Zhang et al. used the improved bat algorithm (IBA) to optimize the parameters of the SVM model and obtained the optimal IBA-SVM diagnosis model to determine the fault type of the rolling bearing[151].

Regularization techniques

Regularization techniques are methods employed during the training of machine learning models to prevent overfitting, improve generalization, and enhance the model's ability to perform well on unseen data. Regularization methods introduce constraints or penalties into the training process, discouraging the model from becoming overly complex and fitting noise in the training data. In the context of training optimization for IFD models, regularization techniques play a crucial role in ensuring that the trained models generalize well to real-world scenarios. Researchers have used these techniques with various models for their improvement. Wen et al. introduced an improved snapshot ensemble learning with diversity regulation[152]. The method generated and combined diverse local minima using cyclical learning rate and diversity regularization to form an ensemble model. Chen et al. proposed a method to diagnose faults in rotating machinery under different working conditions using transfer learning and manifold regularization[153]. The method used the adaptation regularization based on transfer learning (ARTL) framework, which consisted of three steps: (1) domain adaptation using joint distribution, (2) manifold regularization using local neighborhood, and (3) classifier construction using kernel function and structural risk minimization. Qin et al. proposed a novel fault diagnosis method based on enhanced multi-scale sample entropies and balanced adaptation regularization-based transfer learning[154]. The method can deal with the problem of inconsistent data distribution between different working conditions of rotating machinery.

Incremental learning

Chuan Li et al. have presented a unified framework that integrates four major functions, such as feature extraction, anomaly detection, novelty detection, and recursive knowledge increment[130]. The framework has a novel method for incremental fault diagnosis of bearings, which can learn from initially normal data and gradually detect new fault classes over time. It uses contrastive learning to extract homologous and interclass features of bearings, which can represent the bearing condition modes and improve the diagnosis performance. In a similar type of research, Wang et al. have come up with a manifold-contrastive broad learning system (MCBLS)

that can handle the class imbalance problem of small samples using online updating and active learning[131]. The method constructs a one-class broad-learning classifier based on an inherency-guided contrastive mechanism that preserves the data's inherent structure and diversity. The paper introduces a contrastive strategy to fuse the manifold in the broad learning system, which creates two matrices: the manifold matrix and the manifold-contrastive matrix. These matrices are used to maintain the local invariance and enlarge the gaps among different classes of data. A minimum-error strategy based on active learning to annotate the unlabeled samples by comparing the error-pair values of the classifiers is proposed as well. The authors have applied an incremental learning strategy to update the broad-learning classifier online by absorbing the newly annotated data and embedding the manifold-contrastive matrix in the output weight updating to retain the original data structure and improve the model accuracy. Pengfei Chen et al. propose a new method called Contrastive Cluster Center (CCC) that combines graph and contrastive learning to reduce the subdomain gap between different operating conditions of bearings[124]. Traditional classification assumes a closed-set scenario where all classes are known and the model is trained to recognize them. In contrast, open-set classification deals with situations where the model encounters classes during testing that were not present in the training set. Hongchun Sun et al. proposes an intelligent open-set fault diagnosis method for rolling bearings based on the integration of prototype and reconstructed networks. They aim to solve the problem of misjudging unknown faults as specific types of known defects[129].

Data augmentation

Data augmentation is a technique commonly used in tuning optimization, especially in the context of deep learning and machine learning models. While data augmentation is typically associated with image data, similar principles can be applied to other types of data. Some of the data augmentation methods used in tuning optimization are time series data augmentation, feature level augmentation, mixup, cutmix, SMOTE (synthetic minority over-sampling technique) and random erasing. Wang et al. have proposed a novel data augmentation approach with compressed sensing for bearing fault diagnosis. The method generates new and diverse data from limited fault data using compressed sensing theory and a deep convolutional neural network[79]. The authors have provided experimental validation and comparison with other methods using two different bearing datasets. Lyu et al. proposed a novel method to augment time series data for data-driven fault diagnosis in smart manufacturing, which is a challenge due to the scarcity of fault data[155]. The authors introduced a model-independent data augmentation method that combines two existing methods, Gaussian noise and signal stretching, with different weights to generate synthetic fault data. The experimentation reported that the proposed method can significantly improve the fault diagnosis accuracy, and demonstrates its effectiveness and robustness for data-driven fault diagnosis in smart manufacturing. Shi et al. proposed a novel data augmentation framework, combining multibody dynamic simulation (MBS) and fast weighted feature-space averaging (FWFSA), to improve the performance of machine learning-based machine fault diagnosis (MFD) models under various operating conditions[156]. The framework can generate reality-augmented simulation faulty data (RASF), which extends the training data distribution and enhances the model's robustness against condition variations, such as speed, load, sensor position, and interference.

Transfer learning

Transfer learning methods can be employed in the context of tuning optimization to leverage knowledge gained from previous tasks or domains. There are several techniques in the literature which are used in the context. First is pre-trained models, the knowledge gained from the pre-trained model, including tuned hyperparameters, can be transferred to the new optimization task, providing a starting point for further tuning. Second is hyper-parameter transfer, which is found effective in a similar context are known, they can be transferred or used as initial values for hyperparameter optimization in a new task. Other techniques include meta-learning, knowledge distillation, and ensemble methods. Ai et al. have proposed a fully simulated-data-driven transfer-learning method that uses a domain-invariant data-transform method to convert domain-variant datasets to domain-invariant datasets. The method is based on hidden Markov model (HMM) and deep learning and does not require real data in the training process[157]. Su et al. have proposed a new method called data reconstruction hierarchical recurrent meta-learning (DRHRML) for bearing fault diagnosis with small samples under different working conditions[158]. In the meta-learning stage, a recurrent meta-learning algorithm (RML) is proposed to classify bearing faults using one-shot learning. RML can learn from different tasks and adapt to new tasks quickly and effectively. Ruiyi et al. introduced the design and implementation of three modules that perform feature extraction, feature fusion, and distance measurement for fault diagnosis, using multiscale convolution, dilated convolution, stochastic pooling, and relation network[159]. Authors have adopted a meta-learning strategy to train the model parameters through multiple different tasks, which enables the model to quickly adapt to new faults with few labeled samples. Zhang et al. have proposed a novel method that uses a self-attention mechanism and transfer learning to diagnose and classify multiple bearing faults in various motor drive systems[58].

Role of remaining useful life (RUL) prediction in IMHM

Overview

Remaining Useful Life refers to the estimated time a system or component has left before it becomes unable to perform its intended function. In the context of machinery, equipment, or systems, RUL prediction is a crucial aspect of prognostics, which involves forecasting the future health and performance of the system. RUL estimation helps in assessing when maintenance or replacement should be performed to ensure the reliability and availability of the system. RUL estimation allows for early warnings about potential faults or degradation in the system. By predicting how much useful life remains, maintenance actions can be planned proactively, reducing the risk of unexpected failures and minimizing downtime. Accurately predicting RUL is a complex

problem that involves calculations of health indicators making it a tough domain to be applied in real-time scenarios. It is a complete domain to be reviewed, however, in this survey, we covered research literature for resource-restrained environments especially focusing on real-time scenarios.

RUL prediction techniques

The prediction of Remaining Useful Life (RUL) plays a crucial role in intelligent fault diagnosis by providing proactive insights into the future operational state of a system or component. There are several ways in which RUL prediction contributes to intelligent fault diagnosis, proactive maintenance, resource optimization, improved system reliability, reduced downtime, condition-based monitoring (CbM) enhancement, and dynamic operating conditions compatibility[172]. Alfarizi et al. proposed a data-driven framework that uses empirical mode decomposition, random forest, and Bayesian optimization to estimate the failure time of rolling bearings under unknown working conditions[94]. The framework consists of two phases: feature extraction and RUL prediction. In the first phase, they use empirical mode decomposition to decompose the vibrational data into different frequency bands and extract statistical features from each band. In the second phase, they use random forest as a regression model to predict the RUL of bearings and optimize its hyperparameters using Bayesian optimization. Run-to-failure data is used in this experiment for validation of the proposed methodology. Jin et al. have developed a data-driven approach to estimate and predict health conditions and the remaining useful life of bearings based on self-organizing maps and feature fusion[171]. Haitao Wang et al. have proposed a method based on a combination of convolutional attention mechanism and temporal convolutional network[172]. The authors have designed an efficient adaptive shrinkage model that can eliminate noise interference and improve the accuracy of RUL prediction. Xu et al. propose a new hybrid model based on extendable useful life (EUL) under continuous monitoring and bearing status classification. The model does not seek the remaining useful life (RUL) of the bearings but determines if the useful life can be extended to the next maintenance cycle[173]. The authors analyze the statistical properties of typical time domain features extracted from vibration data. They evaluate the correlation of these features with bearing status and select the most sensitive and robust ones for prognostics. The framework proposed in Ref.[174] consists of two phases: feature extraction and RUL prediction. A convolutional autoencoder and an attention-based bidirectional GRU are used as the feature extractor and the predictor, respectively. Kumar et al. propose a novel framework that integrates degradation monitoring, defect identification, and remaining useful life estimation for bearings[95]. The framework leverages a new directed divergence measure to compare probability distributions and generate a health indicator that captures the health status of the bearings. Gibran et al. proposed a data-driven framework for predicting the remaining useful life (RUL) of bearings using empirical mode decomposition, random forest, and Bayesian optimization[175]. Then random forest technique integrates many predictors (regression trees) generated randomly using the bootstrapping method. The hyperparameters of the random forest are tuned by Bayesian optimization to improve the RUL prediction accuracy.

Cloud-edge enabled real-time fault diagnosis in IMHM

Overview

As sensor technology advances persistently, the cost of individual sensors continues to decline. Consequently, an increasing number of sensors are being installed in machinery and equipment. Given this scenario, cloud computing emerges as a fitting paradigm for handling extensive volumes of big data. Extensive research and application efforts have been devoted to exploring cloud computing in the realms of machine condition monitoring and fault diagnosis. Explorations into cloud and edge computing within the domains of machine data processing and fault diagnosis have yielded remarkable and encouraging outcomes[184]. One of the main challenges in this domain is the use of modern artificial intelligence efficiently to achieve real-time monitoring and maintenance. The recent techniques in deep learning especially neural networks have demonstrated notable success, offering a comprehensive solution for diagnosing bearing faults. By stacking network depth and utilizing search techniques in high-dimensional space, CNNs aim to enhance performance. Despite the impressive achievements of deep neural networks (DNNs), which come at the cost of substantial computing and storage overhead, certain challenges persist in the implementation of these models[185]. Fog computing plays a significant role in intelligent fault diagnosis by providing a decentralized and distributed computing infrastructure that complements edge and cloud computing. The fog computing layer can perform initial data processing and analysis locally. This is particularly beneficial for filtering and preprocessing data before sending relevant information to the cloud. It helps in optimizing bandwidth usage and reducing the load on the central cloud. Teoh et al. presented research on IoT and fog-computing-based predictive maintenance model for asset management in Industry 4.0 using machine learning. It proposes a genetic algorithm (GA) for resource scheduling and a two-class logistic regression for equipment failure prediction[176]. The system architecture consists of five layers: asset, perception, network, fog computing, and cloud computing. The asset layer includes physical, virtual, and human assets. The perception layer collects data from sensors. The network layer transmits data to fog and cloud servers. The fog computing layer enables low-latency and real-time applications. The cloud computing layer handles resource management, big data, and machine learning tasks. Asutkar et al. have presented a Tiny ML framework based on transfer learning and a lightweight convolutional neural network (CNN) for vibration-based fault diagnosis of different machines[50]. The authors demonstrate the feasibility of their approach for edge inference and online training on Raspberry Pi and ESP32 microcontroller boards. They show that their model has only 2498 parameters and can achieve high accuracy and low latency for fault classification.

Data compression can be a valuable technique in the context of Intelligent Fault Diagnosis (IFD) in cloud-edge implementations, addressing certain challenges associated with data transmission and storage. Data compression can be beneficial in certain ways. It can reduce bandwidth usage, improve transmission speeds, lower storage requirements, minimize latency, and improve privacy and security. Yin et al. have proposed a novel

method that reduces the size of vibration data while preserving the fault information. The method compresses the data in three dimensions: length, width, and height[186].

Frameworks

The importance of making deep learning models more interpretable and explainable for bearing fault diagnosis is discussed in Ref.[68]. Authors have explored the possibilities and benefits of designing lightweight and efficient deep learning models for bearing fault diagnosis, especially for edge computing scenarios. It compares some existing methods of reducing the model size and complexity, such as distillation, pruning, and sparse networks, and shows their performance on bearing fault datasets. Fu et al. have developed a novel system (EdgeCog) for real-time bearing fault diagnosis based on lightweight edge computing and deep learning. It uses a microcontroller to run a compressed CNN model with an attention mechanism to analyze vibration data and identify faults[185]. The system uses a technique to reduce the size and complexity of the CNN model by using quantization, pruning, and distillation methods. It aims to make the model suitable for resource-constrained edge devices. Zhang et al. have presented a framework called DeepHealth, which consists of two submodels: DH-1 for health perception and DH-2 for sequence prediction. They use an enhanced attention mechanism to capture global dependencies from vibrational data and leverage the long- and short-term sequence prediction of sensor data to support instant maintenance decision-making. They also conduct a destructive experiment on a real IIoT-enabled rotating machinery and construct a balanced industrial dataset for model evaluations[60]. He et al. have proposed a real-time fault diagnosis framework based on sound signal analysis, improved cyclostationary analysis, and edge computing. The method can diagnose motor bearing faults in a non-contact way and display the results on an LCD screen[177]. The framework is based on an experimental device and hardware system, which consists of BLDCM, a microphone sensor, a hall signal output port with a micro-controller unit (MCU). A real-time display of the diagnosis results is shown on a liquid crystal display (LCD) screen connected to the edge computing system. Liu et al. has utilized edge-cloud architecture to develop an intelligent digital-twin prediction and reverse control system[179]. The system uses a long short-term memory (LSTM) based error prediction model. Wang et al. proposed a novel method for applying federated learning to industrial fault diagnosis using a multi-branch neural network (MBNN) model that can reduce computation and communication costs for edge nodes[187]. The model is composed of multiple branches that can handle different types of faults, and how the edge nodes can obtain and update only the relevant branches from the cloud. It also proposes an asynchronous federated optimization algorithm that can adjust the update weights of nodes based on their training accuracy. Billel Bengherbia et al. has presented an FPGA-based edge device for data analysis and fault detection[181]. The solution converts acquired data into two-dimensional feature representations, which are then passed to a simplified CNN for fault predictions. In a similar research deep learning methods are used on edge-cloud collaborative architecture to analyse multimedia data[183]. Vidushi Goyal et al. explore the domain of user-specific models on mobile platforms[188]. The authors have developed a hardware-friendly, lightweight pruning technique to create user-specific models on mobile platforms, while also performing inference. The research focuses on the idea of creating smaller, customized machine-learning models for edge devices, rather than using generic, compute-intensive models that cater to a diverse range of users. Wan et al. propose a semi-supervised graph neural network method for intelligent fault diagnosis of bearings using multi-sensor data[178]. The method leverages global-local mutual information maximization within the unsupervised encoder and global-global information maximization between the supervised and unsupervised encoders. This approach effectively mitigates the requirement for extensive and expensive labeling of monitoring data for the purpose of fault diagnosis.

Limitations and challenges

Intelligent systems in industrial process control face several challenges including (1) data quality and availability issues, (2) feature extraction and selection, (3) interpretability, in deep learning models it becomes difficult to understand the decision rationale (4) generalization, the models trained in one domain usually don't perform at the same level when applied to slightly different domain[222], (5) class imbalance, (6) real-time decision making (7) integration. The literature surveyed covers the research on all these topics and still there is a lot of room for improvement to overcome these challenges.

Cross-machine knowledge generalization refers to the ability of machine learning models to apply the knowledge learned from one or multiple domains (the "seen" domains) to a new, previously unseen domain1. This is of great importance for developing and deploying machine learning applications in real-world conditions.

Continuous degradation mode diagnosis, In many real-world systems, degradation is a continuous process. For example, a machine part does not go from being in perfect condition to being completely broken instantly. Instead, it degrades over time, with its performance gradually worsening. This degradation can be due to various factors such as wear and tear, aging, environmental conditions, etc.[189]. Continuous degradation mode diagnosis aims to track this gradual degradation process and diagnose faults at different stages of degradation. This allows for more timely and accurate fault diagnosis, which can help in taking appropriate preventive measures and reducing downtime.

The decision-making process of CNNs demands substantial storage and computing power. However, in real-world scenarios, especially for embedded systems, the extensive framework of CNN faces constraints such as installation size, energy, network support, and computing resources. Consequently, there is a pressing need for edge computing systems employing lightweight neural network models for fault diagnosis. In essence, the primary challenge lies in implementing real-time diagnostic methods on resource-constrained embedded systems. In the domain of intelligent fault diagnosis for embedded systems, there is a noticeable absence of research studies focusing on critical features associated with fault mechanisms. While many CNNs have proven

successful in fault diagnosis, assessing the effectiveness of CNN decision-making based on temporal attention maps poses a challenge.

In situations where network conditions and task dynamics undergo rapid changes, the effectiveness of data-driven intelligent algorithms is hindered. This challenge arises because these algorithms struggle to acquire thorough statistics for precise predictions, leading to a decline in the performance of computational offloading and complicating adaptive adjustments[180]. Enhancing environment-aware, intelligent optimization poses a current challenge. The aim is to enable the computational offloading algorithm to dynamically adjust to changes in network conditions and task requirements, ultimately achieving global multi-objective optimization.

Conclusion and future prospective

The technology, size, and cost of data acquisition devices have improved over the years enabling deeper integration down the industrial domain on scalar level. The online distributed processing capabilities of the cloud have encouraged micro-services to be used for data processing. The third element of this trilogy is the advancement in the domain of artificial intelligence. This trilogy has played a pivotal role in the current industrial revolution. In the past few years, theoretical investigations into intelligent fault diagnosis have produced significant advancements. Nonetheless, intelligent diagnostic models encounter challenges in practical engineering scenarios, including issues like data scarcity, data imbalance, noise, real-time decision support, and variable operating conditions.

Failure prevention depends upon the early fault detection that may lead to the failure. The ideal and desired responsiveness is real-time. However, balancing accuracy with responsiveness is a key challenge. Relevant data is collected through sensors capturing information on performance and anomalies that is communicated between cloud and edge devices for real-time data exchange. To minimize data transfer to the cloud critical information is processed at the edge (initial data analysis performed locally on edge), reducing latency in fault detection. Cloud utilizes deep learning models for comprehensive fault diagnosis with historical data fused with edge-processed data. Incremental learning is applied to incorporate the evolving behavior of the system.

There are several promising problems in data processing, raw data often contains noise and anomalies that can mislead diagnostic models. Incomplete or missing data can hinder the training of accurate models. Outliers may distort the representation of normal system behavior. Researchers have used various techniques to improve data quality ranging from interpolation, transformation, normalization, and standardization. The research[7] has utilized a third-order spline curve interpolation for data enhancement. In fault diagnosis, the occurrence of faults might be significantly less frequent than normal conditions, leading to imbalanced datasets. Imbalanced datasets can bias the learning process, making the model less adept at identifying faults. Researchers have worked with different techniques like resampling, synthetic data generation, transfer learning, and ensemble methods. One of the main challenges is the fusion of data from multiple heterogeneous sensors[28]. Identifying and selecting relevant features for fault diagnosis can be challenging. Ensuring that data is normalized and scaled appropriately for different algorithms and models can be a complex task. One of the commonly used techniques in the context of time-series data or sequential data is random window sampling (RWS)[70]. This approach enables the acquisition of the necessary number of samples from the original data without being constrained by the length of the original dataset. It provides a means to extract the desired quantity of samples, overcoming limitations associated with the dataset's size or length. This method ensures flexibility in generating a sufficient number of samples, unhindered by the constraints imposed by the inherent length of the initial dataset. Addressing the challenges posed by a large influx of Industrial Internet of Things (IIoT) data from numerous sensors across hierarchical and multidimensional sources, the creation of an IIoT-driven intelligent manufacturing system introduces complexities related to efficient data ingestion and management, precise analysis, and prediction of equipment conditions. Research directions include handling partially labeled or unlabeled sensor data, managing substantial computational loads, achieving real-time streaming data processing, and prompt response mechanisms[48].

The promising research direction in model construction with deep learning is the availability of an inadequate amount of labeled training data, particularly for rare or complex faults, which can hinder model construction. Limited representation of diverse fault scenarios may lead to biased models. Skewed Class Ratios and imbalanced datasets, where normal conditions significantly outnumber faults, may result in models biased toward normal behavior. Relevant Feature Identification, Identifying the most relevant features for fault diagnosis can be challenging, particularly in complex systems. High-Dimensional Data, dealing with high-dimensional data and selecting informative features is a non-trivial task. Transfer learning challenges, transferring knowledge from one domain to another may not always yield optimal results. Generalization across systems, difficulty in developing models that generalize well across different types of systems or industries. Model drift, ongoing monitoring is essential to detect and address model drift, ensuring models remain accurate over time. Update procedures, establishing procedures for updating models with new data or adapting to changing conditions.

Training optimization methods have certain research gaps like class imbalance, a scarcity of labeled data for certain fault classes may hinder the training of models, leading to incomplete representations and unequal distribution of fault classes may bias models towards the majority class, limiting the ability to detect minority faults. Noisy labels, and inaccurate or noisy labels in the training data can misguide the learning process and a lack of information for all relevant features may impede model training. In real operating environments conditions are harsh and lead to noise[2]. Hybrid research based on signal processing and deep learning techniques can lead to solutions with better noise handling and optimized training. Hyperparameter tuning, a lot of research has been put into methods for tuning hyperparameters, which can be challenging. This task can be computationally expensive and time-consuming therefore balancing out the performance is crucial. Smart manufacturing systems aim for automated modeling in dynamic industrial environments. While deep transfer learning has shown success in cross-domain fault diagnosis, many existing DTL algorithms are limited by their dataset and domain

specificity[169]. These algorithms often necessitate hyperparameter optimization (HPO) with prior knowledge to achieve optimal prediction performance. Overfitting and underfitting, a part of the research in predictive health maintenance focuses on establishing a technique that can balance and ensure optimal generalization on unseen data. To enhance generalization, researchers in Ref.[27] have used a dropout layer between two classification layers with a configurable ratio. Adaptation delays and changing conditions over time may lead to concept drift, requiring models to adapt to evolving fault patterns. Delays in adapting to dynamic environments may result in reduced model accuracy. One of the major practical requirements of modern predictive health maintenance systems is a labeled dataset[130]. The researchers have focused on fault diagnosis augmented by active learning and data annotation with small and unbalanced samples[131].

Over the past couple of years, the field of fault diagnosis has seen the introduction of intelligent techniques, with a substantial volume of research papers being published on the subject. This review systematically examines and discusses recent publications, categorizing intelligent fault diagnosis methods into three distinct groups: data processing methods, model construction methods, and training optimization methods.

Data availability

The datasets analysed during the current study are available in the CWRU Bearing Dataset repository, https://www.kaggle.com/datasets/brjapon/cwru-bearing-datasets.

Received: 26 June 2024; Accepted: 6 November 2024
Published online: 07 January 2025

References

1. Zhang, X., Zhao, B. & Lin, Y. Machine learning based bearing fault diagnosis using the case western reserve university data: A review. *IEEE Access* **9**, 155598–155608 (2021).
2. Meng, Z., Cui, Z., Liu, J., Li, J. & Fan, F. Maximum cyclic gini index deconvolution for rolling bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
3. Tang, X., Xu, Z. & Wang, Z. A novel fault diagnosis method of rolling bearing based on integrated vision transformer model. *Sensors* **22**(10), 3878 (2022).
4. Shenfield, A. & Howarth, M. A novel deep learning model for the detection and identification of rolling element-bearing faults. *Sensors* **20**(18), 5112 (2020).
5. Qi, B., Li, Y., Yao, W. & Li, Z. Application of emd combined with deep learning and knowledge graph in bearing fault. *J. Signal Process. Syst.* **1**, 1–20 (2023).
6. Jin, Y., Hou, L. & Chen, Y. A time series transformer based method for the rotating machinery fault diagnosis. *Neurocomputing* **494**, 379–395 (2022).
7. Han, T., Pang, J. & Tan, A. C. Remaining useful life prediction of bearing based on stacked autoencoder and recurrent neural network. *J. Manuf. Syst.* **61**, 576–591 (2021).
8. Zhang, J., Chen, J., Deng, H. & Hu, W. A novel framework based on adaptive multi-task learning for bearing fault diagnosis. *Energy Rep.* **9**, 522–531 (2023).
9. Ghorvei, M., Kavianpour, M., Beheshti, M. T. & Ramezani, A. Synthetic to real framework based on convolutional multi-head attention and hybrid domain alignment. In *2022 8th International Conference on Control, Instrumentation and Automation (ICCIA)* 1–6 (IEEE, 2022).
10. Rajput, D. S., Meena, G., Acharya, M. & Mohbey, K. K. Fault prediction using fuzzy convolution neural network on iot environment with heterogeneous sensing data fusion. *Meas. Sens.* **26**, 100701 (2023).
11. Hou, Y., Wang, J., Chen, Z., Ma, J. & Li, T. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer. *Eng. Appl. Artif. Intell.* **124**, 106507 (2023).
12. Yang, D., Karimi, H. R. & Gelman, L. An explainable intelligence fault diagnosis framework for rotating machinery. *Neurocomputing* **541**, 126257 (2023).
13. Magar, R., Ghule, L., Li, J., Zhao, Y. & Farimani, A. B. Faultnet: A deep convolutional neural network for bearing fault classification. *IEEE Access* **9**, 25189–25199 (2021).
14. Wang, H., Zhang, W., Yang, D. & Xiang, Y. Deep-learning-enabled predictive maintenance in industrial internet of things: Methods, applications, and challenges. *IEEE Syst. J.* **17**, 2602 (2022).
15. Alonso-González, M. et al. Bearing fault diagnosis with envelope analysis and machine learning approaches using cwru dataset. *IEEE Access* **11**, 57796 (2023).
16. Tang, L., Wu, X., Wang, D. & Liu, X. A comparative experimental study of vibration and acoustic emission on fault diagnosis of low-speed bearing. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
17. Cateni, S. et al. Variable selection through genetic algorithms for classification purposes. In *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications*, vol. 1, 6–11 (AIA, 2010)
18. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection—A review and recommendations for the practicing statistician. *Biom. J.* **60**, 431. https://doi.org/10.1002/bimj.201700067 (2018).
19. Tang, G., Hu, H., Kong, J. & Liu, H. A novel fault feature selection and diagnosis method for rotating machinery with symmetrized dot pattern representation. *IEEE Sens. J.* **23**(2), 1447–1461 (2022).
20. Lee, C.-Y., Le, T.-A. & Hung, C.-L. A feature selection approach based on memory space computation genetic algorithm applied in bearing fault diagnosis model. *IEEE Access* **11**, 51282 (2023).
21. Yang, Y., Liu, H., Han, L. & Gao, P. A feature extraction method using vmd and improved envelope spectrum entropy for rolling bearing fault diagnosis. *IEEE Sens. J.* **23**(4), 3848–3858 (2023).
22. Gu, J., Peng, Y., Lu, H., Chang, X. & Chen, G. A novel fault diagnosis method of rotating machinery via vmd, cwt and improved cnn. *Measurement* **200**, 111635 (2022).
23. Zhao, Y., Zhang, N., Zhang, Z. & Xu, X. Bearing fault diagnosis based on mel frequency cepstrum coefficient and deformable space-frequency attention network. *IEEE Access* **11**, 34407–34420 (2023).
24. Zhou, C. et al. A mechanical part fault diagnosis method based on improved multiscale weighted permutation entropy and multiclass lstsvm. *Measurement* **214**, 112671 (2023).
25. Kulevome, D. K. B., Wang, H. & Wang, X. Rolling bearing fault diagnostics based on improved data augmentation and convnet. *J. Syst. Eng. Electron.* **34**(4), 1074–1084 (2023).
26. Liu, X., Sun, W., Li, H., Wang, Z. & Li, Q. Imbalanced sample fault diagnosis of rolling bearing using deep condition multidomain generative adversarial network. *IEEE Sens. J.* **23**(2), 1271–1285 (2022).
27. Huo, J., Qi, C., Li, C. & Wang, N. Data augmentation fault diagnosis method based on residual mixed self-attention for rolling bearings under imbalanced samples. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).

28. Tong, J., Liu, C., Bao, J., Pan, H. & Zheng, J. A novel ensemble learning-based multisensor information fusion method for rolling bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1–12 (2022).
29. Zhou, H. et al. Hob vibration signal denoising and effective features enhancing using improved complete ensemble empirical mode decomposition with adaptive noise and fuzzy rough sets. *Expert Syst. Appl.* **233**, 120989 (2023).
30. Xiong, J. et al. A bearing fault diagnosis method based on improved mutual dimensionless and deep learning. *IEEE Sens. J.* **23**(16), 18338 (2023).
31. Yu, W., Pi, D., Xie, L. & Luo, Y. Multiscale attentional residual neural network framework for remaining useful life prediction of bearings. *Measurement* **177**, 109310 (2021).
32. Hosna, A. et al. Transfer learning: A friendly introduction. *J. Big Data* **9**(1), 102 (2022).
33. Zhu, W., Shi, B., Feng, Z. & Tang, J. An unsupervised domain adaptation method for intelligent bearing fault diagnosis based on signal reconstruction by cycle-consistent adversarial learning. *IEEE Sens. J.* **1**, 1 (2023).
34. Zhu, W., Shi, B. & Feng, Z. A transfer learning method using high-quality pseudo labels for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1–11 (2022).
35. Yu, X. et al. A wavelet packet transform-based deep feature transfer learning method for bearing fault diagnosis under different working conditions. *Measurement* **201**, 111597 (2022).
36. Ayodeji, A. et al. Causal augmented convnet: A temporal memory dilated convolution model for long-sequence time series prediction. *ISA Trans.* **123**, 200–217 (2022).
37. Liu, S., Chen, J., He, S., Shi, Z. & Zhou, Z. Few-shot learning under domain shift: Attentional contrastive calibrated transformer of time series for fault diagnosis under sharp speed variation. *Mech. Syst. Signal Process.* **189**, 110071 (2023).
38. Yu, X. et al. A new cross-domain bearing fault diagnosis framework based on transferable features and manifold embedded discriminative distribution adaption under class imbalance. *IEEE Sens. J.* **23**(7), 7525–7545 (2023).
39. Gao, H., Zhang, X., Gao, X., Li, F. & Han, H. Icot-gan: Integrated convolutional transformer gan for rolling bearings fault diagnosis under limited data condition. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
40. Luo, P., Yin, Z., Yuan, D., Gao, F. & Liu, J. An intelligent method for early motor bearing fault diagnosis based on Wasserstein distance generative adversarial networks meta learning. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
41. Ren, Z., Ji, J., Zhu, Y., Hong, J. & Feng, K. Generative adversarial network with dual multi-scale feature fusion for data augmentation in fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
42. Lu, Z., Cai, Z., Qian, W. & Zhou, D. Intelligent fault diagnosis of bearings with both working condition variation and target data scarcity. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
43. Azari, M. S., Flammini, F., Santini, S. & Caporuscio, M. A systematic literature review on transfer learning for predictive maintenance in industry 4.0. *IEEE Access* **11**, 12887 (2023).
44. Castano, F., Cruz, Y. J., Villalonga, A. & Haber, R. E. Data-driven insights on time-to-failure of electromechanical manufacturing devices: A procedure and case study. *IEEE Trans. Ind. Inform.* **19**, 7190 (2022).
45. Mao, W., Chen, J., Liu, J. & Liang, X. Self-supervised deep domain-adversarial regression adaptation for online remaining useful life prediction of rolling bearing under unknown working condition. *IEEE Trans. Ind. Inf.* **19**(2), 1227–1237 (2022).
46. Ni, Q., Ji, J. & Feng, K. Data-driven prognostic scheme for bearings based on a novel health indicator and gated recurrent unit network. *IEEE Trans. Ind. Inf.* **19**(2), 1301–1311 (2022).
47. Gao, H., Li, Y., Zhao, Y. & Song, Y. Dual channel feature-attention-based approach for rul prediction considering the spatiotemporal difference of multisensor data. *IEEE Sens. J.* **23**, 8514 (2023).
48. Yu, W., Liu, Y., Dillon, T. & Rahayu, W. Edge computing-assisted iot framework with an autoencoder for fault detection in manufacturing predictive maintenance. *IEEE Trans. Ind. Inf.* **19**(4), 5701–5710 (2022).
49. Zhao, C., Tang, B., Huang, Y. & Deng, L. Edge collaborative compressed sensing in wireless sensor networks for mechanical vibration monitoring. *IEEE Trans. Ind. Inform.* **19**, 8852 (2022).
50. Asutkar, S., Chalke, C., Shivgan, K. & Tallur, S. Tinyml-enabled edge implementation of transfer learning framework for domain generalization in machine fault diagnosis. *Expert Syst. Appl.* **213**, 119016 (2023).
51. Kamath, V. & Renuka, A. Deep learning based object detection for resource constrained devices: Systematic review, future trends and challenges ahead. *Neurocomputing* **531**, 34–60. https://doi.org/10.1016/j.neucom.2023.02.006 (2023).
52. Gutierrez-Torre, A. et al. Automatic distributed deep learning using resource-constrained edge devices. *IEEE Internet Things J.* **9**(16), 15018–15029. https://doi.org/10.1109/JIOT.2021.3098973 (2022).
53. Ren, Z. et al. A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
54. Zhang, Q., Yuan, R., Lv, Y., Li, Z. & Wu, H. Multivariate dynamic mode decomposition and its application to bearing fault diagnosis. *IEEE Sens. J.* **23**(7), 7514–7524 (2023).
55. Niu, G., Liu, E., Wang, X., Ziehl, P. & Zhang, B. Enhanced discriminate feature learning deep residual cnn for multitask bearing fault diagnosis with information fusion. *IEEE Trans. Ind. Inf.* **19**(1), 762–770 (2022).
56. Brusamarello, B., Silva, J. C. C., Morais Sousa, K. & Guarneri, G. A. Bearing fault detection in three-phase induction motors using support vector machine and fiber Bragg grating. *IEEE Sens. J.* **23**(5), 4413–4421 (2022).
57. Liu, D., Cui, L. & Cheng, W. Flexible generalized demodulation for intelligent bearing fault diagnosis under nonstationary conditions. *IEEE Trans. Ind. Inf.* **19**(3), 2717–2728 (2022).
58. Zhang, X. et al. Inferable deep distilled attention network for diagnosing multiple motor bearing faults. *IEEE Trans. Transp. Electrif.* **9**, 2207 (2022).
59. Kim, T. & Lee, S. A novel unsupervised clustering and domain adaptation framework for rotating machinery fault diagnosis. *IEEE Trans. Ind. Inform.* **19**, 9404 (2022).
60. Zhang, W. et al. Deephealth: A self-attention based method for instant intelligent predictive maintenance in industrial internet of things. *IEEE Trans. Ind. Inf.* **17**(8), 5461–5473 (2020).
61. Meng, Z., Zhu, J., Cao, S., Li, P. & Xu, C. Bearing fault diagnosis under multi-sensor fusion based on modal analysis and graph attention network. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
62. Chang, M., Yao, D. & Yang, J. Intelligent fault dignosis of rolling bearings using efficient and lightweight resnet networks based on an attention mechanism. *IEEE Sens. J.* **23**, 9136 (2023).
63. Xue, L., Lei, C., Jiao, M., Shi, J. & Li, J. Rolling bearing fault diagnosis method based on self-calibrated coordinate attention mechanism and multi-scale convolutional neural network under small samples. *IEEE Sens. J.* **23**, 10206 (2023).
64. Wang, H. et al. Fault diagnosis method for imbalanced data of rotating machinery based on time domain signal prediction and sc-resnest. *IEEE Access* **11**, 38875 (2023).
65. Wang, D., Li, Y., Jia, L., Song, Y. & Wen, T. Attention-based bilinear feature fusion method for bearing fault diagnosis. *IEEE/ASME Trans. Mechatron.* **28**, 1695 (2022).
66. Wang, X., Zhang, H. & Du, Z. Multi-scale noise reduction attention network for aero-engine bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
67. Mao, W., Liu, K., Zhang, Y., Liang, X. & Wang, Z. Self-supervised deep tensor domain-adversarial regression adaptation for online remaining useful life prediction across machines. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
68. Pu, H., Zhang, K. & An, Y. Restricted sparse networks for rolling bearing fault diagnosis. *IEEE Trans. Ind. Inform.* **19**, 11139 (2023).

69. Wan, S. et al. Bearing fault diagnosis based on multi-sensor information coupling and attentional feature fusion. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
70. Meng, Z., Luo, C., Li, J., Cao, L. & Fan, F. Research on fault diagnosis of rolling bearing based on lightweight model with multiscale features. *IEEE Sens. J.* **23**, 13236 (2023).
71. Lee, C.-Y. & Zhuo, G.-L. Identifying bearing faults using multiscale residual attention and multichannel neural network. *IEEE Access* **11**, 26953–26963 (2023).
72. Ma, W., Zhang, Y., Ma, L., Liu, R. & Yan, S. An unsupervised domain adaptation approach with enhanced transferability and discriminability for bearing fault diagnosis under few-shot samples. *Expert Syst. Appl.* **225**, 120084 (2023).
73. Yan, X., Zhang, C.-A. & Liu, Y. Multi-branch convolutional neural network with generalized shaft orbit for fault diagnosis of active magnetic bearing-rotor system. *Measurement* **171**, 108778 (2021).
74. Buchaiah, S. & Shakya, P. Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection. *Measurement* **188**, 110506. https://doi.org/10.1016/j.measurement.2021.110506 (2022).
75. Yang, K., Zhao, L. & Wang, C. A new intelligent bearing fault diagnosis model based on triplet network and svm. *Sci. Rep.* **12**, 5234 (2022).
76. Shao, H. et al. Dual-threshold attention-guided gan and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation. *IEEE Trans. Ind. Inform.* **19**, 9933 (2023).
77. Li, J. et al. A new probability guided domain adversarial network for bearing fault diagnosis. *IEEE Sens. J.* **23**(2), 1462–1470 (2022).
78. Han, B., Jiang, X., Wang, J. & Zhang, Z. A novel domain adaptive fault diagnosis method for bearings based on unbalance data generation. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
79. Wang, D., Dong, Y., Wang, H. & Tang, G. Limited fault data augmentation with compressed sensing for bearing fault diagnosis. *IEEE Sens. J.* **23**(13), 14499 (2023).
80. Ren, H., Wang, J., Shen, C., Huang, W. & Zhu, Z. Dual classifier-discriminator adversarial networks for open set fault diagnosis of train bearings. *IEEE Sens. J.* **1**, 1 (2023).
81. Su, Z. et al. Cross-domain open-set fault diagnosis based on target domain slanted adversarial network for rotating machinery. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
82. Liu, S., Jiang, H., Wu, Z., Liu, Y. & Zhu, K. Machine fault diagnosis with small sample based on variational information constrained generative adversarial network. *Adv. Eng. Inform.* **54**, 101762 (2022).
83. Dai, Z., Zhao, L., Wang, K. & Zhou, Y. Generative adversarial network to alleviate information insufficiency in intelligent fault diagnosis by generating continuations of signals. *Appl. Soft Comput.* **147**, 110784 (2023).
84. Chen, Q. et al. A lightweight and robust model for engineering cross-domain fault diagnosis via feature fusion-based unsupervised adversarial learning. *Measurement* **205**, 112139 (2022).
85. Li, J., Liu, Y. & Li, Q. Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method. *Measurement* **189**, 110500. https://doi.org/10.1016/j.measurement.2021.110500 (2022).
86. Zhang, J., Zhang, K., An, Y., Luo, H. & Yin, S. An integrated multitasking intelligent bearing fault diagnosis scheme based on representation learning under imbalanced sample condition. *IEEE Trans. Neural Netw. Learn. Syst.* **1**, 1–12 (2023).
87. Liu, X. et al. Cross-domain intelligent bearing fault diagnosis under class imbalanced samples via transfer residual network augmented with explicit weight self-assignment strategy based on meta data. *Knowl. Based Syst.* **251**, 109272. https://doi.org/10.1016/j.knosys.2022.109272 (2022).
88. Ding, Y. et al. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. *Reliab. Eng. Syst. Saf.* **230**, 108890. https://doi.org/10.1016/j.ress.2022.108890 (2023).
89. Wang, X., Jiang, H., Liu, Y., Liu, S. & Yang, Q. A dynamic spectrum loss generative adversarial network for intelligent fault diagnosis with imbalanced data. *Eng. Appl. Artif. Intell.* **126**, 106872. https://doi.org/10.1016/j.engappai.2023.106872 (2023).
90. Liu, X. et al. A fault diagnosis method of rolling bearing based on improved recurrence plot and convolutional neural network. *IEEE Sens. J.* **23**, 10767 (2023).
91. Yuan, Z., Ma, Z., Li, X. & Li, J. A multichannel mn-gcn for wheelset-bearing system fault diagnosis. *IEEE Sens. J.* **23**(3), 2481–2494 (2022).
92. Lyu, P., Zhang, K., Yu, W., Wang, B. & Liu, C. A novel rsg-based intelligent bearing fault diagnosis method for motors in high-noise industrial environment. *Adv. Eng. Inform.* **52**, 101564. https://doi.org/10.1016/j.aei.2022.101564 (2022).
93. Liang, P. et al. Intelligent fault diagnosis of rolling bearing based on wavelet transform and improved resnet under noisy labels and environment. *Eng. Appl. Artif. Intell.* **115**, 105269. https://doi.org/10.1016/j.engappai.2022.105269 (2022).
94. Alfarizi, M. G., Tajiani, B., Vatn, J. & Yin, S. Optimized random forest model for remaining useful life prediction of experimental bearings. *IEEE Trans. Ind. Inform.* **19**, 7771 (2022).
95. Kumar, A., Parkash, C., Tang, H. & Xiang, J. Intelligent framework for degradation monitoring, defect identification and estimation of remaining useful life (rul) of bearing. *Adv. Eng. Inform.* **58**, 102206 (2023).
96. Hua, L., Wu, X., Liu, T. & Li, S. The methodology of modified frequency band envelope kurtosis for bearing fault diagnosis. *IEEE Trans. Ind. Inf.* **19**(3), 2856–2865 (2022).
97. Li, Y., Zhou, J., Li, H., Meng, G. & Bian, J. A fast and adaptive empirical mode decomposition method and its application in rolling bearing fault diagnosis. *IEEE Sens. J.* **23**(1), 567–576 (2022).
98. Chen, Z. et al. Feature extraction based on hierarchical improved envelope spectrum entropy for rolling bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
99. Zhou, Q., Yi, C., Yan, L., Huang, C. & Lin, J. A blind deconvolution approach based on spectral harmonics-to-noise ratio for rotating machinery condition monitoring. *IEEE Trans. Autom. Sci. Eng.* **20**(2), 1092–1107 (2022).
100. Chen, R., Huang, Y., Xu, X., Zhang, X. & Qiu, T. Rolling bearing fault feature extraction method using adaptive maximum cyclostationarity blind deconvolution. *IEEE Sens. J.* **23**, 17761 (2023).
101. Li, J., Liu, Y. & Xiang, J. Optimal maximum cyclostationary blind deconvolution for bearing fault detection. *IEEE Sens. J.* **23**, 15975 (2023).
102. Yi, C. et al. An adaptive harmonic product spectrum for rotating machinery fault diagnosis. *IEEE Trans. Instrum. Meas.* **72**, 1–12 (2022).
103. Pan, H., Xu, H. & Zheng, J. A novel symplectic relevance matrix machine method for intelligent fault diagnosis of roller bearing. *Expert Syst. Appl.* **192**, 116400 (2022).
104. Ma, C., Yang, Z., Zhang, K., Xiang, L. & Xu, Y. Optimization of Ramanujan subspace periodic and its application in identifying industrial bearing fault features. *IEEE Trans. Instrum. Meas.* **72**, 1–7 (2022).
105. Mitra, S. & Koley, C. Early and intelligent bearing fault detection using adaptive superlets. *IEEE Sens. J.* **23**(7), 7992–8000 (2023).
106. Zhao, H. et al. Intelligent diagnosis using continuous wavelet transform and gauss convolutional deep belief network. *IEEE Trans. Reliab.* **72**, 692 (2022).
107. Xue, Y., Yang, R., Chen, X., Tian, Z. & Wang, Z. A novel local binary temporal convolutional neural network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
108. Cui, X. et al. A novel fault diagnosis method for rotor-bearing system based on instantaneous orbit fusion feature image and deep convolutional neural network. *IEEE/ASME Trans. Mechatron.* **28**(2), 1013–1024 (2022).
109. Zhang, B., Pang, X., Zhao, P. & Lu, K. A new method based on encoding data probability density and convolutional neural network for rotating machinery fault diagnosis. *IEEE Access* **11**, 26099–26113 (2023).

110. Li, Q. et al. Fault diagnosis of bearings and gears based on litenet with feature aggregation. *IEEE Trans. Instrum. Meas.* **72**, 1–9 (2023).
111. Liang, H., Cao, J. & Zhao, X. Multibranch and multiscale dynamic convolutional network for small sample fault diagnosis of rotating machinery. *IEEE Sens. J.* **23**(8), 8973–8988 (2023).
112. Liu, X., Lu, J. & Li, Z. Multi-scale fusion attention convolutional neural network for fault diagnosis of aero-engine rolling bearing. *IEEE Sens. J.* **1**, 1 (2023).
113. Cheng, L. et al. S3m: Two-stage-based semi-self-supervised method for intelligent bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
114. Tang, H. et al. Feature extraction of multi-sensors for early bearing fault diagnosis using deep learning based on minimum unscented kalman filter. *Eng. Appl. Artif. Intell.* **127**, 107138 (2024).
115. Huo, C. et al. A class-level matching unsupervised transfer learning network for rolling bearing fault diagnosis under various working conditions. *Appl. Soft Comput.* **146**, 110739 (2023).
116. Li, F., Wang, L., Wang, D., Wu, J. & Zhao, H. An adaptive multiscale fully convolutional network for bearing fault diagnosis under noisy environments. *Measurement* **216**, 112993 (2023).
117. Huo, C., Jiang, Q., Shen, Y., Zhu, Q. & Zhang, Q. Enhanced transfer learning method for rolling bearing fault diagnosis based on linear superposition network. *Eng. Appl. Artif. Intell.* **121**, 105970 (2023).
118. Zhao, X. et al. Multiscale deep graph convolutional networks for intelligent fault diagnosis of rotor-bearing system under fluctuating working conditions. *IEEE Trans. Ind. Inf.* **19**(1), 166–176 (2022).
119. Feng, K. et al. Digital twin enabled domain adversarial graph networks for bearing fault diagnosis. *IEEE Trans. Ind. Cyber Phys. Syst.* **1**, 1 (2023).
120. Lu, F., Tong, Q., Feng, Z. & Wan, Q. Unbalanced bearing fault diagnosis under various speeds based on spectrum alignment and deep transfer convolution neural network. *IEEE Trans. Ind. Inform.* **19**, 8295 (2022).
121. Yang, S., Cui, Z. & Gu, X. A balanced deep transfer network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
122. Li, X., Hu, H., Zhang, S. & Tang, G. A fault diagnosis method for rotating machinery with semi-supervised graph convolutional network and images converted from vibration signals. *IEEE Sens. J.* **23**, 11946 (2023).
123. Yin, P. et al. A multi-scale graph convolutional neural network framework for fault diagnosis of rolling bearing. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
124. Chen, P., Zhao, R., He, T., Wei, K. & Yuan, J. Unsupervised structure subdomain adaptation based the contrastive cluster center for bearing fault diagnosis. *Eng. Appl. Artif. Intell.* **122**, 106141 (2023).
125. Zhu, J. et al. Application of recurrent neural network to mechanical fault diagnosis: A review. *J. Mech. Sci. Technol.* **36**(2), 527–542 (2022).
126. Imamura, L., Avila, S., Pacheco, F., Salles, M. & Jablon, L. Diagnosis of unbalance in lightweight rotating machines using a recurrent neural network suitable for an edge-computing framework. *J. Control Autom. Electr. Syst.* **33**(4), 1272–1285 (2022).
127. Chang, Y., Chen, J., Lv, H. & Liu, S. Heterogeneous bi-directional recurrent neural network combining fusion health indicator for predictive analytics of rotating machinery. *ISA Trans.* **122**, 409–423. https://doi.org/10.1016/j.isatra.2021.04.024 (2022).
128. Zhang, Z. et al. Attention gate guided multiscale recursive fusion strategy for deep neural network-based fault diagnosis. *Eng. Appl. Artif. Intell.* **126**, 107052. https://doi.org/10.1016/j.engappai.2023.107052 (2023).
129. Sun, H., Yang, B. & Lin, S. An open set diagnosis method for rolling bearing faults based on prototype and reconstructed integrated network. *IEEE Trans. Instrum. Meas.* **72**, 1–10 (2022).
130. Li, C. et al. Incrementally contrastive learning of homologous and interclass features for the fault diagnosis of rolling element bearings. *IEEE Trans. Ind. Inform.* **19**, 11182 (2023).
131. Wang, N. et al. Manifold-contrastive broad learning system for wheelset bearing fault diagnosis. *IEEE Trans. Intell. Transp. Syst.* **24**, 9886 (2023).
132. Zhu, Z. et al. A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. *Measurement* **1**, 112346 (2022).
133. Xu, Z. et al. Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors. *Renew. Energy* **182**, 615–626. https://doi.org/10.1016/j.renene.2021.10.024 (2022).
134. Shi, J. et al. Planetary gearbox fault diagnosis using bidirectional-convolutional lstm networks. *Mech. Syst. Signal Process.* **162**, 107996. https://doi.org/10.1016/j.ymssp.2021.107996 (2022).
135. An, Y., Zhang, K., Liu, Q., Chai, Y. & Huang, X. Rolling bearing fault diagnosis method base on periodic sparse attention and lstm. *IEEE Sens. J.* **22**(12), 12044–12053. https://doi.org/10.1109/JSEN.2022.3173446 (2022).
136. Zhi Tang, X. L., Bo, Lin & Wei, D. A semi-supervised transferable lstm with feature evaluation for fault diagnosis of rotating machinery. *Appl. Intell.* **52**, 1703–1717. https://doi.org/10.1007/s10489-021-02504-1 (2022).
137. Zhu, S. et al. A transformer model with enhanced feature learning and its application in rotating machinery diagnosis. *ISA Trans.* **133**, 1–12 (2023).
138. Xu, P. & Zhang, L. A fault diagnosis method for rolling bearing based on 1d-vit model. *IEEE Access* **11**, 39664 (2023).
139. Fang, H. et al. A lightweight transformer with strong robustness application in portable bearing fault diagnosis. *IEEE Sens. J.* **23**, 9649 (2023).
140. Wu, H., Triebe, M. J. & Sutherland, J. W. A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: A rotary system application. *J. Manuf. Syst.* **67**, 439–452 (2023).
141. Sun, Z., Wang, Y. & Gao, J. Intelligent fault diagnosis of rotating machinery under varying working conditions with global-local neighborhood and sparse graphs embedding deep regularized autoencoder. *Eng. Appl. Artif. Intell.* **124**, 106590 (2023).
142. Shi, M. et al. Deep hypergraph autoencoder embedding: An efficient intelligent approach for rotating machinery fault diagnosis. *Knowl. Based Syst.* **260**, 110172 (2023).
143. Shi, M. et al. Graph embedding deep broad learning system for data imbalance fault diagnosis of rotating machinery. *Reliab. Eng. Syst. Saf.* **240**, 109601 (2023).
144. Chen, X., Guo, Y. & Na, J. Instantaneous-angular-speed-based synchronous averaging tool for bearing outer race fault diagnosis. *IEEE Trans. Ind. Electron.* **70**(6), 6250–6260 (2022).
145. Gwak, M., Kim, M. S., Yun, J. P. & Park, P. Robust and explainable fault diagnosis with power-perturbation-based decision boundary analysis of deep learning models. *IEEE Trans. Ind. Inform.* **19**, 6982 (2022).
146. Chen, C., Shi, J., Shen, M., Feng, L. & Tao, G. A predictive maintenance strategy using deep learning quantile regression and kernel density estimation for failure prediction. *IEEE Trans. Instrum. Meas.* **72**, 1–12 (2023).
147. Hongwei, F., Ceyi, X., Jiateng, M., Xiangang, C. & Xuhui, Z. A novel intelligent diagnosis method of rolling bearing and rotor composite faults based on vibration signal-to-image mapping and cnn-svm. *Meas. Sci. Technol.* **34**(4), 044008. https://doi.org/10.1088/1361-6501/acad90 (2023).
148. Lee, S. & Kim, T. Impact of deep learning optimizers and hyperparameter tuning on the performance of bearing fault diagnosis. *IEEE Access* **11**, 55046–55070. https://doi.org/10.1109/ACCESS.2023.3281910 (2023).
149. Ye, X., Gao, L., Li, X. & Wen, L. A new hyper-parameter optimization method for machine learning in fault classification. *Appl. Intell.* **53**(11), 14182–14200 (2023).
150. Zhang, Y., Liu, W., Wang, X. & Shaheer, M. A. A novel hierarchical hyper-parameter search algorithm based on greedy strategy for wind turbine fault diagnosis. *Expert Syst. Appl.* **202**, 117473. https://doi.org/10.1016/j.eswa.2022.117473 (2022).

151. Zhang, M., Yin, J. & Chen, W. Rolling bearing fault diagnosis based on time-frequency feature extraction and iba-svm. *IEEE Access* **10**, 85641–85654. https://doi.org/10.1109/ACCESS.2022.3198701 (2022).
152. Wen, L., Xie, X., Li, X. & Gao, L. A new ensemble convolutional neural network with diversity regularization for fault diagnosis. *J. Manuf. Syst.* **62**, 964–971. https://doi.org/10.1016/j.jmsy.2020.12.002 (2022).
153. Chen, R., Zhu, Y., Yang, L., Hu, X. & Chen, G. Adaptation regularization based on transfer learning for fault diagnosis of rotating machinery under multiple operating conditions. *IEEE Sens. J.* **22**(11), 10655–10662. https://doi.org/10.1109/JSEN.2022.3165398 (2022).
154. Hu, Q., Si, X., Qin, A., Lv, Y. & Liu, M. Balanced adaptation regularization based transfer learning for unsupervised cross-domain fault diagnosis. *IEEE Sens. J.* **22**(12), 12139–12151. https://doi.org/10.1109/JSEN.2022.3174396 (2022).
155. Lyu, P., Zhang, H., Yu, W. & Liu, C. A novel model-independent data augmentation method for fault diagnosis in smart manufacturing. In *Leading Manufacturing Systems Transformation—Proceedings of the 55th CIRP Conference on Manufacturing Systems* 949–954 (2022).
156. Shi, D., Ye, Y., Gillwald, M. & Hecht, M. Robustness enhancement of machine fault diagnostic models for railway applications through data augmentation. *Mech. Syst. Signal Process.* **164**, 108217. https://doi.org/10.1016/j.ymssp.2021.108217 (2022).
157. Ai, T. et al. Fully simulated-data-driven transfer-learning method for rolling-bearing-fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
158. Su, H., Xiang, L., Hu, A., Xu, Y. & Yang, X. A novel method based on meta-learning for bearing fault diagnosis with small sample learning under different working conditions. *Mech. Syst. Signal Process.* **169**, 108765. https://doi.org/10.1016/j.ymssp.2021.108765 (2022).
159. Ma, R., Han, T. & Lei, W. Cross-domain meta learning fault diagnosis based on multi-scale dilated convolution and adaptive relation module. *Knowl. Based Syst.* **261**, 110175. https://doi.org/10.1016/j.knosys.2022.110175 (2023).
160. Qian, Q., Zhou, J. & Qin, Y. Relationship transfer domain generalization network for rotating machinery fault diagnosis under different working conditions. *IEEE Trans. Ind. Inform.* **19**, 9898 (2023).
161. Fang, H., Liu, H., Wang, X., Deng, J. & An, J. The method based on clustering for unknown failure diagnosis of rolling bearings. *IEEE Trans. Instrum. Meas.* **72**, 1–8 (2023).
162. Liu, X., Sun, W., Li, H., Li, Q. & Lv, S. A fusing domain feature and sharing label space based fault diagnosis approach for different distribution and unlabeled rolling bearing sample. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
163. Liu, Y. et al. A lifelong learning method based on generative feature replay for bearing diagnosis with incremental fault types. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
164. Yue, K., Li, J., Chen, Z., Huang, R. & Li, W. Multiple source-free domain adaptation network based on knowledge distillation for machinery fault diagnosis. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
165. Li, Y., Dong, Y., Xu, M., Liu, P. & Wang, R. Instance weighting based partial domain adaptation for intelligent fault diagnosis of rotating machinery. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
166. Ma, W., Liu, R., Guo, J., Wang, Z. & Ma, L. A collaborative central domain adaptation approach with multi-order graph embedding for bearing fault diagnosis under few-shot samples. *Appl. Soft Comput.* **140**, 110243 (2023).
167. Gao, Q., Huang, T., Zhao, K., Shao, H. & Jin, B. Multi-source weighted source-free domain transfer method for rotating machinery fault diagnosis. *Expert Syst. Appl.* **237**, 121585 (2024).
168. Jiang, Y., Xia, T., Wang, D., Zhang, K. & Xi, L. Joint adaptive transfer learning network for cross-domain fault diagnosis based on multi-layer feature fusion. *Neurocomputing* **487**, 228–242 (2022).
169. Liu, G., Shen, W., Gao, L. & Kusiak, A. Automated broad transfer learning for cross-domain fault diagnosis. *J. Manuf. Syst.* **66**, 27–41 (2023).
170. Li, W., Shang, Z., Gao, M., Liu, F. & Liu, H. Intelligent fault diagnosis of partial deep transfer based on multi-representation structural intraclass compact and double-aligned domain adaptation. *Mech. Syst. Signal Process.* **197**, 110412 (2023).
171. Jin, X., Que, Z., Sun, Y., Guo, Y. & Qiao, W. A data-driven approach for bearing fault prognostics. *IEEE Trans. Ind. Appl.* **55**(4), 3394–3401 (2019).
172. Wang, H., Yang, J., Wang, R. & Shi, L. Remaining useful life prediction of bearings based on convolution attention mechanism and temporal convolution network. *IEEE Access* **11**, 24407–24419 (2023).
173. Xu, G., Hou, D., Qi, H. & Bo, L. High-speed train wheel set bearing fault diagnosis and prognostics: A new prognostic model based on extendable useful life. *Mech. Syst. Signal Process.* **146**, 107050 (2021).
174. Qin, Y. et al. Dynamic weighted federated remaining useful life prediction approach for rotating machinery. *Mech. Syst. Signal Process.* **202**, 110688 (2023).
175. Alfarizi, M. G., Tajiani, B., Vatn, J. & Yin, S. Optimized random forest model for remaining useful life prediction of experimental bearings. *IEEE Trans. Ind. Inf.* **19**(6), 7771–7779. https://doi.org/10.1109/TII.2022.3206339 (2023).
176. Teoh, Y. K., Gill, S. S. & Parlikad, A. K. Iot and fog computing based predictive maintenance model for effective asset management in industry 4.0 using machine learning. *IEEE Internet Things J.* **10**, 2087 (2021).
177. He, C. et al. Real-time fault diagnosis of motor bearing via improved cyclostationary analysis implemented onto edge computing system. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
178. Wan, W., Chen, J. & Xie, J. Mim-graph: A multi-sensor network approach for fault diagnosis of hsr bogie bearings at the iot edge via mutual information maximization. *ISA Trans.* **139**, 574 (2023).
179. Liu, J., Ma, C., Gui, H. & Wang, S. Intelligent digital-twin prediction and reverse control system architecture for thermal errors enabled by deep learning and cloud-edge computing. *Expert Syst. Appl.* **225**, 120122 (2023).
180. Zhu, X. et al. Deep reinforcement learning-based edge computing offloading algorithm for software-defined iot. *Comput. Netw.* **235**, 110006 (2023).
181. Bengherbia, B. et al. Design and hardware implementation of an intelligent industrial iot edge device for bearing monitoring and fault diagnosis. *Arab. J. Sci. Eng.* **1**, 1–17 (2023).
182. Maurya, M., Panigrahi, I., Dash, D. & Malla, C. Intelligent fault diagnostic system for rotating machinery based on iot with cloud computing and artificial intelligence techniques: A review. *Soft Comput.* **1**, 1–18 (2023).
183. Nan, Y., Jiang, S. & Li, M. Large-scale video analytics with cloud-edge collaborative continuous learning. *ACM Trans. Sens. Netw.* **20**, 1 (2023).
184. Lu, S., Lu, J., An, K., Wang, X. & He, Q. Edge computing on iot for machine signal processing and fault diagnosis: A review. *IEEE Internet Things J.* **10**, 11093 (2023).
185. Fu, L. et al. Edgecog: A real-time bearing fault diagnosis system based on lightweight edge computing. *IEEE Trans. Instrum. Meas.* **1**, 1 (2023).
186. Yin, Y., Liu, Z., Zuo, M., Zhou, Z. & Zhang, J. A three-dimensional vibration data compression method for rolling bearing condition monitoring. *IEEE Trans. Instrum. Meas.* **72**, 1–10 (2023).
187. Qizhao, W., Li, Q., Wang, K., Wang, H. & Peng, Z. Efficient federated learning for fault diagnosis in industrial cloud-edge computing. *Comput. Arch. Inform. Numer. Comput.* **103**(10), 2319–2337 (2021).
188. Goyal, V., Das, R. & Bertacco, V. Hardware-friendly user-specific machine learning for edge devices. *ACM Trans. Embedded Comput. Syst.* **21**(5), 1–29 (2022).
189. Li, J., Wang, Y., Zi, Y., Zhang, H. & Wan, Z. Causal disentanglement: A generalized bearing fault diagnostic framework in continuous degradation mode. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 6250 (2021).

190. Li, H., Liu, T., Wu, X. & Li, S. Correlated svd and its application in bearing fault diagnosis. *IEEE Trans. Neural Netw. Learn. Syst.* **1**, 1 (2021).
191. Chen, Z. et al. Explainable deep ensemble model for bearing fault diagnosis under variable conditions. *IEEE Sens. J.* **23**, 17737 (2023).
192. Choudhary, A., Mian, T., Fatima, S. & Panigrahi, B. Passive thermography based bearing fault diagnosis using transfer learning with varying working conditions. *IEEE Sens. J.* **23**(5), 4628–4637 (2022).
193. Yu, X. et al. An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks. *IEEE Trans. Instrum. Meas.* **72**, 1–14 (2023).
194. Zhou, Y., Dong, Y. & Tang, G. Time-varying online transfer learning for intelligent bearing fault diagnosis with incomplete unlabeled target data. *IEEE Trans. Ind. Inform.* **19**, 7733 (2022).
195. Ruan, D., Han, J., Yan, J. & Gühmann, C. Light convolutional neural network by neural architecture search and model pruning for bearing fault diagnosis and remaining useful life prediction. *Sci. Rep.* **13**(1), 5484 (2023).
196. Pilarski, S., Staniszewski, M., Bryan, M., Villeneuve, F. & Varró, D. Predictions-on-chip: model-based training and automated deployment of machine learning models at runtime: For multi-disciplinary design and operation of gas turbines. *Softw. Syst. Model.* **20**, 685–709 (2021).
197. Zhang, W., Chen, D., Xiao, Y. & Yin, H. Semi-supervised contrast learning based on multi-scale attention and multi-target contrast learning for bearing fault diagnosis. *IEEE Trans. Ind. Inform.* **19**, 10056 (2023).
198. Chen, X. et al. Deep transfer learning for bearing fault diagnosis: A systematic review since 2016. *IEEE Trans. Instrum. Meas.* **72**, 1 (2023).
199. Elsamanty, M., Ibrahim, A. & Salman, W. S. Principal component analysis approach for detecting faults in rotary machines based on vibrational and electrical fused data. *Mech. Syst. Signal Process.* **200**, 110559 (2023).
200. Peng, C., Ouyang, Y., Gui, W., Li, C. & Tang, Z. A multi-indicator fusion-based approach for fault feature selection and classification of rolling bearings. *IEEE Trans. Ind. Inform.* **19**, 8635 (2022).
201. Chen, Z., Wu, J., Deng, C., Wang, X. & Wang, Y. Deep attention relation network: A zero-shot learning method for bearing fault diagnosis under unknown domains. *IEEE Trans. Reliab.* **72**(1), 79–89 (2022).
202. Yu, G. et al. Few-shot fault diagnosis method of rotating machinery using novel mcgm based cnn. *IEEE Trans. Ind. Inform.* **19**, 10944 (2023).
203. Du, W., Hu, P., Wang, H., Gong, X. & Wang, S. Fault diagnosis of rotating machinery based on 1d–2d joint convolution neural network. *IEEE Trans. Ind. Electron.* **70**(5), 5277–5285 (2022).
204. Mario, B., Mezhuyev, V. & Tschandl, M. Predictive maintenance for railway domain: A systematic literature review. *IEEE Eng. Manag. Rev.* **51**, 120 (2023).
205. Alenizi, F. A., Abbasi, S., Mohammed, A. H. & Rahmani, A. M. The artificial intelligence technologies in industry 4.0: A taxonomy, approaches, and future directions. *Comput. Ind. Eng.* **1**, 109662 (2023).
206. Jieyang, P. et al. A systematic review of data-driven approaches to fault diagnosis and early warning. *J. Intell. Manuf.* **1**, 1–28 (2022).
207. Chen, C., Fu, H., Zheng, Y., Tao, F. & Liu, Y. The advance of digital twin for predictive maintenance: The role and function of machine learning. *J. Manuf. Syst.* **71**, 581–594 (2023).
208. CaseWestern Reserve University (CWRU). *Bearing Data Center*. https://engineering.case.edu/bearingdatacenter/download-data-file (Case School of Engineering, CWRU).
209. Lessmeier, C., Kimotho, J. K., Zimmer, D. & Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM Society European Conference*, vol. 3 (2016)
210. Yaguo, L. et al. Xjtu-sy rolling element bearing accelerated life test datasets: A tutorial. *J. Mech. Eng.* **55**(16), 1–6 (2019).
211. Zhang, P. Vibration time-frequency images of planetary gearboxes. *IEEE Dataport* **1**, 1. https://doi.org/10.21227/0zxx-m405 (2022).
212. Lee, J., Qiu, H., Yu, G. & Lin, J. *Bearing Data Set. NASA Prognostics Data Repository*. https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository (2007).
213. Shao, S., McAleer, S., Yan, R. & Baldi, P. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Trans. Ind. Inf.* **15**(4), 2446–2455. https://doi.org/10.1109/TII.2018.2864759 (2019).
214. Liu, S. et al. Bearing fault diagnosis based on improved convolutional deep belief network. *Appl. Sci.* **10**(18), 359. https://doi.org/10.3390/app10186359 (2020).
215. Liu, X., Sun, W., Li, H., Hussain, Z. & Liu, A. The method of rolling bearing fault diagnosis based on multi-domain supervised learning of convolution neural network. *Energies* **15**(13), 614. https://doi.org/10.3390/en15134614 (2022).
216. Nectoux, P. et al. Pronostia: An experimental platform for bearings accelarated life test. In *Proceedings of the IEEE International Conference on Prognostics and Health Management, Denver, CO, USA*, vol. 20 (2012).
217. Zhao, Z. et al. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA Trans.* **107**, 224 (2020).
218. Liu, W., Liu, Y., Li, S. & Chen, W. Adaptive time-reassigned synchrosqueezing transform for bearing fault diagnosis. *IEEE Sens. J.* **23**(8), 8545 (2023).
219. Ren, Z., Jiang, Y., Yang, X., Tang, Y. & Zhang, W. Learnable faster kernel-pca for nonlinear fault detection: Deep autoencoder-based realization. *J. Ind. Inf. Integr.* **40**, 100622 (2024).
220. Jiang, Y., Yin, S. & Kaynak, O. Optimized design of parity relation-based residual generator for fault detection: Data-driven approaches. *IEEE Trans. Ind. Inf.* **17**(2), 1449–1458 (2020).
221. Mahesh, T. et al. Data-driven intelligent condition adaptation of feature extraction for bearing fault detection using deep responsible active learning. *IEEE Access* **1**, 1 (2024).
222. Lei, Y. et al. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **138**, 106587. https://doi.org/10.1016/j.ymssp.2019.106587 (2020).

Author contributions

A.S.K conducted the research and wrote the main manuscript M.U.A and M.A.K have supervised the study. W. J.O has proof read the manuscript and added valuable feedback. S.J and M.U.A reviewed the research approach. A.A has proof read the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S. or W.J.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.