Survey on Foundation Models for Prognostics and Health Management in Industrial Cyber-Physical Systems

Ruonan Liu, *Senior Member, IEEE*, Quanhu Zhang, Te Han, *Member, IEEE*, Boyuan Yang, *Member, IEEE*, Weidong Zhang, *Senior Member, IEEE*, Shen Yin, *Fellow, IEEE*, and Donghua Zhou, *Fellow, IEEE*

***Abstract*—Industrial Cyber-Physical Systems (ICPS) integrating disciplines such as computer science, communication technology, and engineering, have become a crucial component of modern manufacturing and industry. However, ICPS faces numerous challenges during long-term operation, including equipment faults, performance degradation, and security threats, etc. To achieve efficient maintenance and management, prognostics and health management (PHM) has been widely applied in the critical tasks of ICPS such as fault prediction, health monitoring, and maintenance decision-making. The emergence of large-scale foundation models (LFMs) like BERT and GPT marks a significant advancement in artificial intelligence (AI) technology, demonstrating substantial application potential in multiple fields. The accumulation of AI technology, rapid development of LFMs, and the abundance of industrial data and industrial process knowledge provide the foundational conditions for the construction and advancement of industrial LFMs. However, there is currently a lack of consensus on applying LFMs of PHM in ICPS, necessitating a systematic review and roadmap to clarify future development directions. To bridge this gap, this survey provides a comprehensive survey and understanding of the recent advances in LFMs of PHM in ICPS. It provides valuable references for decision makers and researchers in the industry, and helps to further improve the reliability, availability and safety of ICPS.**

***Index Terms*—Industrial cyber-physical systems, prognostics and health management, fault diagnosis, remaining useful life, large-scale foundation models, industrial process knowledge.**

Manuscript received 30 January 2024; revised 30 May 2024; accepted 23 June 2024. Date of publication 10 July 2024; date of current version 23 July 2024. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0119900, in part by the National Natural Science Foundation of China under Grant 62206199 and Grant U2141234, in part by Shanghai Science and Technology Program under Grant 22015810300, in part by Hainan Province Science and Technology Special Fund under Grant ZDYF2024GXJS003, in part by Young Elite Scientist Sponsorship Program under Grant YESS20220409, and in part by Alexander von Humboldt Foundation under Grant 1226831. *(Corresponding authors: Te Han; Boyuan Yang.)*

Ruonan Liu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ruonan.liu@sjtu.edu.cn).

Quanhu Zhang is with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: quanhuzhang @tju.deu.cn).

Te Han is with the Center for Energy and Environmental Policy Research, Beijing Institute of Technology, Beijing 100081, China, also with the School of Management, Beijing Institute of Technology, Beijing 100081, China, and also with the Beijing Laboratory for System Engineering of Carbon Neutrality, Beijing 100081, China (e-mail: hante@bit.edu.cn).

Boyuan Yang and Donghua Zhou are with the Center for Advanced Control and Smart Operations, Nanjing University, Suzhou 215163, China (e-mail: yby@nju.edu.cn; zdh@mail.tsinghua.edu.cn).

Weidong Zhang is with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China, and also with the Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: wdzhang@sjtu.edu.cn).

Shen Yin is with the Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: shen.yin@ntnu.no).

Digital Object Identifier 10.1109/TICPS.2024.3425326

I. Introduction

With the rapid development of science and technology and the continuous progress of intelligent manufacturing, a revolution has been developed in industrial systems to make them more intelligent via the information communication technologies (ICT). With the increasing penetration of the ICT in industrial systems, the transformation of traditional industrial settings into the industrial environment on cyber-physical systems (CPS) pushes the development of industrial cyber-physical systems (ICPS) [1]. ICPS is a system that integrates physical machinery, sensors, actuators and communication network systems in industry [2]. The modern equipment is developing in the direction of large-scale, complexity, and intelligence [3]. Once the mechanical equipment failure, may cause major accidents, will bring huge economic losses. If faults can be detected in time and appropriate decisions can be made, accidents can be avoided to the maximum extent possible. On one hand, prognostics and health management (PHM), which includes condition maintenance, fault diagnosis, trend prediction, and life cycle assessment, provides an effective tool for the safety and reliable operation of ICPS [4]; on the other hand, by leveraging advanced technologies such as data analytics, machine learning, and artificial intelligence, ICPS can monitor, control, and optimize physical processes in real time, which can help to improve the accuracy and efficiency of PHM in return.

Traditionally, the PHM for ICPS can be divided into two ways: model-based methods and data-driven methods. Model-based PHM methods usually need to establish system models of complex ICPS based on physical knowledge and

mathematical foundations, which grow more and more complex with the increase of system integration degree. Data-driven PHM methods can provide the state estimation results via analyzing the sensor data in ICPS without the requirement of prior knowledge and therefore develop rapidly. Especially in recent years, data-driven PHM methods in ICPS have undergone significant development, driven by emerging technologies such as artificial intelligence (AI) and machine learning (ML). Classical deep networks such as convolutional neural networks, auto-encoders, and recurrent neural networks have become effective tools for PHM [5], [6]. However, deep learning-based PHM methods in ICPS still faces many challenges. Firstly, the data in ICPS come from different types of equipment and sensors with different formats and features, and the heterogeneous data from multiple sources can limit the representation capability of model. Secondly, decisions require clear explanations and justifications in ICPS, and thus require the development of interpretable PHM models for fault diagnosis and prediction. In addition, existing methods are limited to designing specific models for specific ICPS scenarios, and the trained models cannot be applied to other tasks or even similar tasks. Therefore, PHM models need to have generalization capabilities to cope with complex ICPS scenarios.

Recently, transformer-based large language models (LLMs) (e.g., GPT-3 [7], T5 [8], BERT [9], etc.) and vision foundation models (VFM) (e.g., ViT [10], CLIP [11], SAM [12], etc.) have achieved outstanding performance in language understanding and vision recognition tasks. The success of ChatGPT proves the effectiveness of large-scale foundation models (LFMs), which triggers the research of large models in various industrial domains, such as the medical large model Med-PaLM [13], Ocean large model AI-GOMS [14], and Geographic large model ERNIE-GeoL [15]. How to leverage cross-domain knowledge in industry to build industrial LFMs has sparked great interest among scholars.

To overcome the limitations of existing PHM methods in ICPS, LFMs offer possible solutions for industrial systems. LFMs are trained on large-scale diverse data, learn complex patterns and relationships without explicit feature engineering, and can be fine-tuned to capture domain-specific knowledge of mechanisms and dynamics. Yan et al. summarised the current state of PHM development and proposed raodmaps of LFMs for PHM [16]. Considering the specificity of ICPS, there is no conclusion on how to build LFMs in ICPS, and there is a lack of systematic literature review. To fill this gap, we explore in this article the potential of LFMs to address the ICPS-PHM challenge, focusing on their scalability, adaptability and performance. Through this survey, we aim to provide a comprehensive understanding of the capabilities and limitations of LFMs in industrial systems. By identifying key research directions and challenges, we hope to stimulate further developments in the field and pave the way for the successful integration of AI technologies into ICPS. The contributions of this paper can be summarised as follows.

1) This survey provides a comprehensive review of key technologies and research advances in PHM for ICPS, and summarizes the industrial process knowledge in ICPS.
2) This survey reviews the main components, research progress and practical applications of the LFMs.
3) Taking into account the actual circumstances in ICPS, this survey systematically analyzes how to establish LFMs of PHM in ICPS, emphasizing that industrial knowledge is the key to improving the interpretability and trustworthiness of the models.
4) This survey analyzes the challenges and possible solutions for LFMs of PHM in ICPS, taking into account the data situation in ICPS and the limitations of AI technology.

The rest of the paper is organized as follows. Section II focuses on the architecture of ICPS and key technologies of PHM. Section III describes to the key components of LFMs and its research advances and applications. Section IV presents the industrial process knowledge in ICPS. Section V systematically describes the research developments in PHM for ICPS and how to implement the LFMs. Section VI provides a comprehensive discussion of opportunities and challenges for LFMs in ICPS. Conclusions are presented in Section VII.

II. Prognostics Health Management for ICPS

ICPS is a networked control system that deeply integrates sensing, computing, control, Internet and physical objects, and it is an important part of the national economic construction. ICPS have complex topologies, diverse abnormal threats, and inefficient system recovery after failures. The establishment of ICPS-PHM can not only significantly improve system reliability, recovery efficiency and security, optimize maintenance cost, but also support intelligent decision-making, which is of great significance to promote the development of ICPS. In this section, we introduce the architecture of ICPS and key technologies of PHM in the era of Industry 4.0.

A. Industrial Cyber-Physical Systems

ICPS integrate physical components for sensing and drive with cyber components for computing and communication to monitor, control, and automate the operation of industrial processes, and typically consist of a physical system and an information system. The physical system includes sensors, controllers, actuators, and corresponding data transmission network components. The information system consists of data transmission network components, data storage components, and control and computing infrastructure components to interconnect, operate, and intelligently manage ICPS. Therefore, the architecture of ICPS is divided into three layers: physical layer, cyber layer, and application layer [17], and data signals are transmitted through various network protocols and gateway components in cyber layer, as shown in Fig. 1. At present, ICPS have been widely used in electric power systems [18], transportation [19], industrial manufacturing [20] and other fields. With the rapid development of technologies such as wireless sensors, cloud computing, and Internet of Things (IoT), the application of ICPS will be even more extensive [21].

When faults occur in ICPS, the physical and chemical production processes can be severely impacted. Due to the

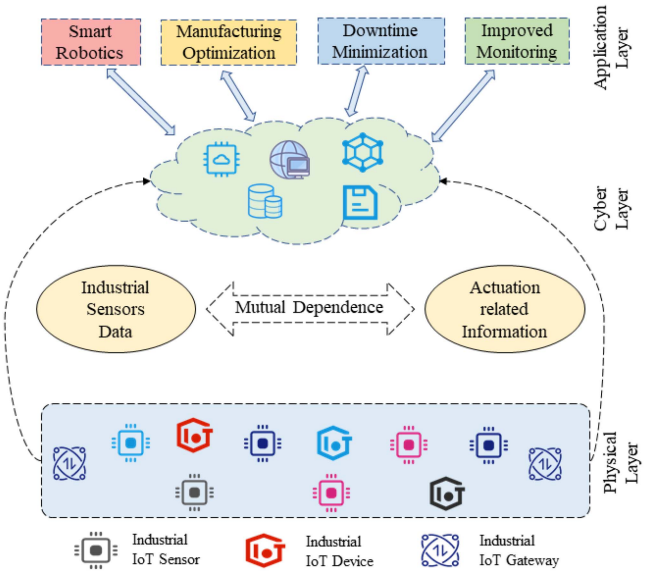

Fig. 1. The general architecture of ICPS.

presence of seepage effects, these effects may propagate rapidly through internal coupling relationships, leading to large-scale cascading failures, which in turn cause massive economic losses, environmental pollution, and even casualties. Therefore, it is necessary to establish an evaluation and certification system for ICPS safety protection and expedite the implementation of ICPS security measures [17]. PHM has become an indispensable maintenance strategy in modern industry through its advantages of advance warning, optimized maintenance, enhanced efficiency and safety, etc. The application of PHM to ICP can effectively prevent and mitigate the impact of faults, ensure the continuity and stability of the production process, reduce economic losses, protect the environment and personnel safety. This holds substantial economic value and social significance.

B. Prognostics Health Management

PHM aims to comprehensively utilize equipment sensor data, expert knowledge, and maintenance support resources, leveraging AI methods and reasoning models to achieve equipment condition monitoring, fault identification and diagnosis, health status assessment and prediction, and ultimately provide maintenance and other health management measures [22]. Previously, ICPS maintenance strategies have evolved through four stages: reactive maintenance, periodic preventive maintenance, condition-based maintenance, and predictive maintenance. PHM technology integrates the concepts of condition-based maintenance and predictive maintenance, successfully achieving effective fault diagnosis and early prevention of equipment faults.

PHM includes data acquisition and transmission of sensor networks, equipment condition monitoring, fault diagnosis, remaining useful life prediction and health management, etc, and the framework as shown in Fig. 2. First, status data such as vibration, rotational speed and temperature are collected by sensors deployed on the equipment. Considering the influence of working conditions and noise, signal preprocessing methods are essential to ensure the quality of data. Subsequently, fault feature information is extracted from data, including signal processing-based methods and deep learning-based methods. Then, anomaly detection, fault diagnosis and prediction are performed based on the fault features. Finally, maintenance decisions and recommendations are made based on the fault diagnosis and prediction results. Fault diagnosis and remaining useful life prediction are key techniques of PHM, which are essential for achieving effective management and maintenance of equipment [23].

1) Fault Diagnosis: The evolution of fault diagnosis has progressed from experience-driven, to model-driven, and then to knowledge-driven and data-driven [24]. Initially, fault diagnosis mainly relies on the experience and intuition of maintenance personnel, which suffered from high subjectivity and low efficiency. Subsequently, fault diagnosis began to introduce mathematical models, establishing physical or mathematical models of the equipment to achieve more objective and precise fault identification. In the 1990 s, knowledge-based fault diagnosis began to emerge. These methods extract expert experience and diagnosis rules to establish fault knowledge base, enabling more intelligent fault diagnosis.

In recent years, with the advancement of sensor technology and Big Data analysis techniques, data-driven diagnosis methods based on massive operational data have gradually matured. The intelligent fault diagnosis (IFD) method based on deep learning can adaptively extract features from vibration signals, providing excellent fault diagnosis results while significantly improving the efficiency of diagnosis. However, due to the complexity of industrial equipment structures and the variability of operating environments, IFD faces various challenges. To ensure stable and reliable fault diagnosis, these challenges must be carefully considered and addressed.

Firstly, equipment monitoring data typically originate from multiple sensors and various measurement points, such as vibration, temperature, and pressure. This multi-source heterogeneous data provides information about equipment health from multiple dimensions but they can impact reduce the model's ability to learn representations. Secondly, industrial equipment usually operates under healthy conditions, so there is much more health data than failure data. Therefore, industrial data suffers from small samples and long-tailed distributions. Therefore, industry suffers from data imbalance problems such as few-shot and long-tailed distributions. Additionally, deep learning models generally assume that training and testing data are independent and identically distributed (i.i.d.), but this assumption may not hold for industry. The operational conditions of the equipment, including speed, load, and environmental factors, are highly complex and variable. As a result, there is distribution drift between training and testing data, which hinders the diagnosis accuracy of deep learning models in real-world scenarios.

To address these practical challenges, researchers have been actively exploring various methods. Currently, the focus of deep learning-based fault diagnosis research includes information fusion [25], data imbalance and few-shot learning [26] as well

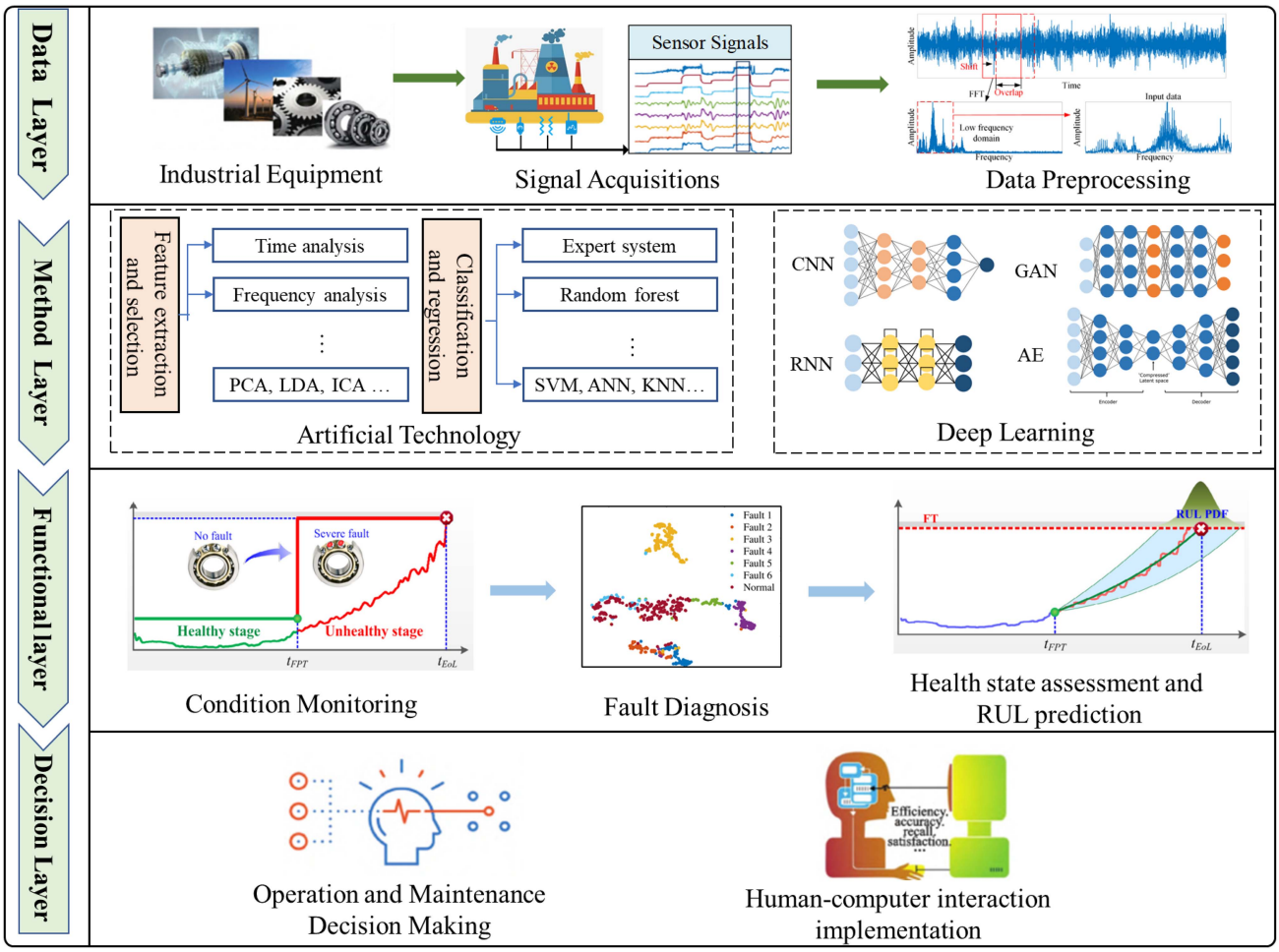

Fig. 2. Illustration of PHM framework of ICPS.

as domain transfer and domain adaptation [27]. These advanced techniques promise to significantly improve the accuracy and reliability of IFD for industrial equipment.

2) Remaining Useful Life Prediction: Fault prediction aims to provide early warnings of potential faults in industrial equipment, predicting degradation trends and remaining useful life (RUL). Based on current classification methods, RUL prediction models are divided into two categories: physics-based models and data-driven models [28]. Physics-based models are constructed through mathematical or physical models of degradation phenomena in industrial system components, using specialized models to characterize degradation and substitute existing data into the models to determine RUL. Data-driven models rely on previously observed data to predict the future state of the system or infer RUL by matching similar historical patterns. Data-driven models can effectively model highly nonlinear, complex, and multidimensional systems without requiring prior expert knowledge of the physical behavior.

The core aspects of RUL include: 1) assessing the current health status of the equipment, constructing health indicators (HI), and analyzing potential degradation trends; 2) estimating the time when the equipment may fail in the future and predicting the RUL. The construction of HI for industrial equipment is fundamental for RUL prediction, as an appropriate HI construction method can ensure the accuracy of subsequent fault predictions. Accurate RUL prediction provides a basis for determining the optimal maintenance time of equipment, thus achieving economic operation and maintenance. Traditional physics-based RUL prediction methods need to consider the internal fault mechanisms of the equipment, which can be limiting when applied to complex industrial equipment. Data-driven methods can directly extract the changing patterns of health status characteristics from condition monitoring data. Data-driven RUL prediction methods within deep learning frameworks have promising application prospects [29], [30].

III. Large-Scale Foundation Models

The parameter sizes and training data volumes of LFMs have rapidly increased, significantly enhancing model performance. Consequently, the research and application of LFMs have become hot topics in artificial intelligence. In this section, we explore the key Components of LFMs and their development and applications.

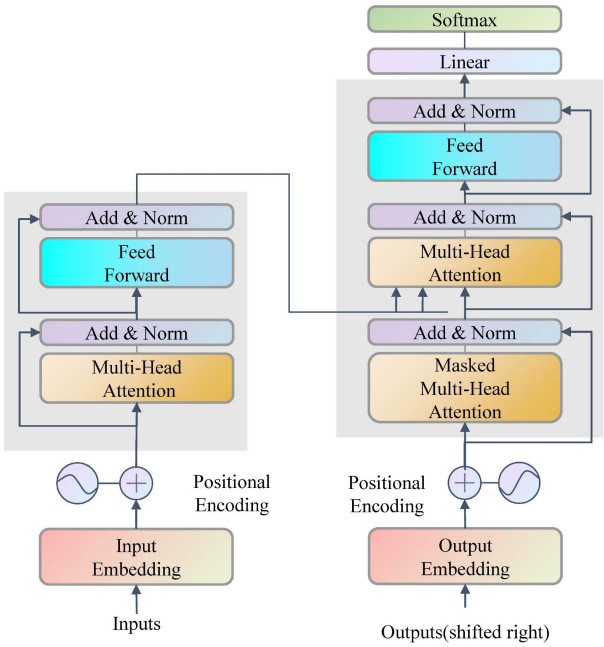

Fig. 3. Model architecture of the transformer.

A. Key Components of LFMs

1) Multiple Self-Attention Mechanism: Self-attention (SA) is the basic module in Transformer. SA projects the input sequences into the set of queries Q, keys K, and values V with dimension C via three learnable linear mapping matrices W_Q, W_K, W_V, and then obtains the self-attention weights by using the following formula:

$$SA(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{1}$$

By linearly transforming the input sequences, SA is able to capture the semantic features and distant dependencies of the input sequences. Multi-headed self-attention (MSA) is an extension of SA, which consists of n SA heads, realizes the attention operation by paralleling it, and splices the outputs of all the SA heads after a linear projection layer:

$$MSA(Q,K,V) = \text{Concat}\left(SA_1, SA_2, L, SA_n\right) * W^0 \tag{2}$$

where W^0 denotes the weight of the fully connected layer for fusing the output weights of multiple attention headers.

2) Transformer: With its exceptional model capacity and parallelization capabilities, the Transformer has become the standard backbone model for developing various LFMs. The architecture of Transformer as depicted in Fig. 3, consists of a stack of multiple encoders and decoders [31]. Each encoder is composed of two basic components: a MSA module and a feed-forward network (FFN) module. The MSA module employs a SA module to learn the relationships within the input sequences, while the FFN module includes an activation function and two linear normalization layers (LayerNorm). The MSA and FFN modules utilize residual connections and layer normalization structures. Given an input x_0 and its positional embedding x_{pos}, the output x_k of the $k-th$ encoder can be represented as:

$$\begin{aligned} x_0 &= x_0 + x_{pos} \\ x_k &= \text{LayerNorm}(x_{k-1} + \text{MSA}(x_{k-1})) \\ x_k &= \text{LayerNorm}(x_k + \text{FFN}(x_k)) \end{aligned} \tag{3}$$

The decoder consists of two MSA modules and one linear layer. The first MSA module adds a one-way attention mask so that the input embedding vectors can only focus on past embedding vectors and themselves, ensuring that the prediction results depend only on the generated output lexical elements. Subsequently, the output of the masked multi-attention module is processed along with the output of the encoder through a second MSA module. The visual Transformer has a similar structure, with the difference that the input is a combination of 2D image embedding vectors, positional coding and category embedding vectors.

B. LFMs for NLP and CV

Transformer effectively addresses the issue of long-term dependencies in long sequence inputs, and its parallelism enhances training efficiency and alleviates problems such as gradient vanishing and exploding due to excessively large models. Transformer has achieved remarkable results in various tasks within natural language processing (NLP) and computer vision (CV), laying a solid foundation for the rapid development of LFMs [32].

In 2018, Google introduced BERT [9], the first LLM with over 300 million parameters. RoBERTa [33] utilized more training data and resources, introduced a dynamic masking strategy that achieved state-of-the-art on multiple tasks. In the same year, OpenAI released GPT-1 [34], which uses autoregressive models for pre-training. Subsequently, GPT-2 [35] increased the number of parameters to 1.5 billion by extending the model capacity and data diversity. In 2020, OpenAI released the GPT-3 [7]. GPT-3 extended the model architecture based on GPT-2, and the number of parameters reached 175 billion, realising a quantum leap in the number of parameters of the model. ChatGPT based on GPT-3 aroused widespread attention to AI in the society. GPT-3.5 [36] utilized comparative learning of text embedding and code embedding to greatly enhance the inference of the model. In 2022, Google released the PaLM [37] with a staggering number of 540 billion parameters. In February 2023, Meta AI launched LLaMA [38], followed by the release of the more powerful LLaMA-2 [39] in September, which greatly advanced the progress of LFMs.

Vision Transformer (ViT) [10] applied Transformer to CV for the first time, validating the feasibility of Transformer as a unified vision model architecture. In 2021, OpenAI proposed the large-scale visual language model CLIP [11], which verified the effectiveness of large-scale weakly supervised pre-training on text-image combination. In 2022, Nanjing University proposed video masked autoencoders (VideoMAE) [40], which extended pre-trained large-scale models to the video field for video tasks such as action recognition and action detection. In 2023, Meta AI proposed the segment anything model (SAM) [12], a generalized

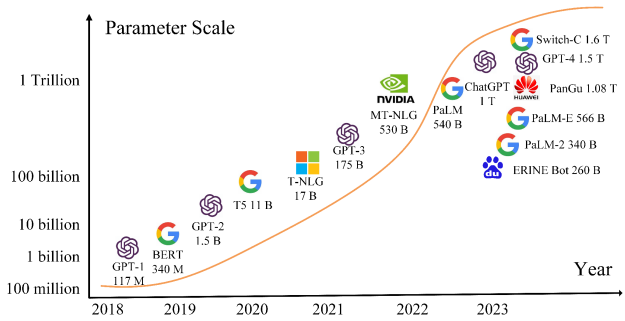

Fig. 4. The trends of the LLM scale changes.

image segmentation model trained on the Segment Anything 1- Billion mask dataset [41]. SAM used prompt engineering for downstream segmentation tasks and is able to generalise to an unprecedented number of new objects without the need to fine-tune downstream tasks. The PaLM-E [42] was the first large-scale multimodal model to handle a variety of of embodied reasoning tasks from a variety of observational modalities on multiple implementations. GPT-4 extended [43] input from textual data to multimodal data. The development of LFMs is shown in Fig. 4.

C. LFMs for Time Series

Time series analysis is crucial in fields such as retail sales forecasting, time series missing value filling, and industrial anomaly detection, etc. Many scholars have explored the possibility of using LFMs for time series analysis. PromptCast [44] is a new paradigm for time series prediction based on LLMs. PromptCast transformed numerical values into prompts and constructed prediction tasks in a sentence-to-sentence manner, enabling the application of LLMs in prediction. TIME-LLM [45] utilized the powerful pattern recognition and inference capabilities of LLMs to match time series data with natural language, demonstrating superior performance in few-shot and zero-shot learning. LLM4TS [46], TEST [47], TEMPO [48], and LLMT-IME [49] are also time series foundation models based on LLMs, contributing to the construction of a unified framework for time series modeling.

TimeGPT [50] is the first time series LFMs to be trained using more than 100 billion data from finance, meteorology, energy, network traffic, etc. and allows users to fine-tune it with their own data, which ultimately supports various forecasting and anomaly detection tasks. TimeGPT opens up new possibilities for time series analysis and is promising to drive further development in related fields.

Considering the characteristics of time series data, many scholars have conducted extensive research on time series LFMs in specific fields. Yu et al. explored the use of LLMs to forecast stock returns in the financial market [51], and Xie et al. conducted a zero-shot analysis of ChatGPT's ability to forecast multimodal stock movements [52]. Brown et al. demonstrated that, with only minor tuning, LLMs can be fundamentally trained on a wide range of physiological and behavioural time-series data and make meaningful inferences about affairs in clinical and health settings [53]. AuxMobLCast made predictions about future human mobility by turning human movement data into natural language sentences and fine-tuning a pre-trained LFMs [54].

D. LFMs for Industry

After going through the stages of mechanization, electrification, automation, and informatization, industrial development is currently transitioning from digitalization to intelligence. A large amount of data, foundational capabilities, and application scenarios have been accumulated in ICPS, providing a solid foundation for the integration of ICPS with AI technologies. AI is gradually demonstrating human-like understanding and analytical capabilities. The integration of these capabilities with industrial scenarios is introducing intelligence into industrial production, maintenance, and health management, potentially advancing ICPS towards an adaptive, self-decision-making, and self-executing intelligent stage.

In November 2023, SmartMore released the first industrial multimodal large model, IndustryGPT V1.0. The training data for IndustryGPT V1.0 includes five major disciplines of optics, mechanics, electronics, computer science, and software engineering, and comprehensive knowledge of industries such as equipment, mining, electric power, petrochemicals, and construction, etc., covering more than 200 different industrial scenarios and over 3 million industrial images. It can understand scene intentions, easily answer questions in production environments, and provide precise decision-making support. COSMO-GPT is released by COSMOPlat in January 2024. Based on an open-source general large model, COSMO-GPT enhances its performance on industrial tasks through knowledge injection, model fusion, and model judgment. With COSMOPlat's technical accumulation in AI and massive industrial data, COSMO-GPT boasts over 10 billion parameters and integrates more than 3,900 mechanistic models and over 200 expert algorithm libraries. Its functionality spans intelligent question answering, text generation, image recognition, database querying, and decision support, etc.

The accumulation of AI technology, the rapid development of large models, and the abundance of industrial data and knowledge provide fundamental conditions for the construction and development of industrial large models. However, existing industrial large models primarily function as question-answering support systems and are not yet capable of directly participating in the process monitoring and health management of ICPS. Leveraging industrial data, knowledge, and existing large model technologies to construct LFMs of PHM can contribute to the stable operation and healthy development of ICPS, holding significant social and economic value.

IV. Knowledge in ICPS

Data-driven ICPS-PHM methods lack interpretability, although they can provide highly accurate prediction and diagnosis results. Enhancing the interpretability can boost the credibility of the model as well as subsequent optimisation of the model,

thereby achieving more efficient and reliable PHM. Industrial systems encompass extensive industrial knowledge, such as equipment structure knowledge, fault mechanism knowledge, and historical knowledge of equipment [55]. In ICPS, the richer the knowledge stored in the system, the better the problem solving ability. Integrating industrial knowledge into industrial foundation models is an effective method to enhancing their stability and interpretability [56].

Fault diagnosis, prediction and processing in ICPS mainly involve three basic elements: diagnosis object, diagnosis system and diagnosis knowledge. Combining the characteristics of ICPS, expert experience and mechanism knowledge, we classify industrial knowledge into four major categories.

A. Equipment Knowledge

Equipment knowledge includes structural, functional and behavioural knowledge. Structural knowledge refers to the components of a equipment or system and their connecting relationships, covering different levels of systems, components and parts. Functional knowledge relates to the specific functions that need to be achieved in the design of equipment or systems, and each structural component fulfill a particular function. The lack of function of any structure and component may directly or indirectly lead to fault. Behavioural knowledge describes the state of a equipment or system, which is reflected by the performance parameters of the system during operation.

B. System Knowledge

System knowledge includes model knowledge and historical knowledge. Model knowledge covers deep understanding of system principles, models, and equations, such as input-output energy transformations and the dynamic characteristics of equipment or systems. Historical knowledge covers the manufacture, installation, monitoring, diagnosis, maintenance records, and fault history of the equipment.

C. Fault Knowledge

Fault knowledge comprises mechanism knowledge, consequence knowledge, processing knowledge and feature knowledge. Mechanism knowledge is obtained by comprehensively analysing the structure, function and behaviour of equipment to obtain the rule of formation and development of equipment failure. After the occurrence of equipment failure, the direct or indirect impact on the components or the whole system is called consequence knowledge. Processing knowledge refers to the measures that should be taken after the occurrence of equipment or system fault. Feature knowledge originates from the relationship between equipment operating characteristics and faults accumulated in practice by experts and operators, and is used to speculate, identify and verify equipment faults.

D. Prior Knowledge

Prior knowledge includes diagnosis standards, prediction standards, information processing knowledge, and auxiliary knowledge. The knowledge of methods, models and criteria used for fault diagnosis is described as diagnostic criteria. Prediction criteria cover models, methods and judgement criteria related to fault prediction. Information processing knowledge involves information acquisition, computation, analysis and feature extraction, such as time series analysis, wavelet analysis and time-frequency analysis. Auxiliary knowledge refers to background and environmental knowledge related to diagnosis and processing, such as fault behaviour of similar equipment, environmental climate conditions and monitoring systems.

V. Large-Scale Foundation Models for Prognostics Health Management in ICPS

Data-driven PHM methods have achieved significant success in ICPS, which utilize advanced data analytics, machine learning, and deep learning techniques to achieve equipment failure prediction, maintenance schedule optimization, and operational efficiency improvement. However, these methods are typically trained and optimized for specific ICPS scenarios and tasks, leading to limitations in generalization, multi-task processing, and cognitive capabilities. In industrial environments, hundreds of core components require health monitoring and fault prediction. It is impractical to develop separate deep models for each device and subsystem. Furthermore, the working mechanisms of data-driven PHM methods often function as “black boxes”, lacking interpretability and transparency, which hinders fault cause analysis and expert decision support.

LFMs have demonstrated outstanding zero-shot generalization and powerful multi-task processing abilities, particularly in time series analysis, which demonstrates strong data processing capability. The success of LFMs offers an effective solution to the above challenges. To promote the research and application of LFMs in the ICPS-PHM field, this section illustrates and analyzes in detail how to construct LFMs of PHM in ICPS applications from four aspects.

A. Large-Scale Datasets in ICPS

Data in the ICPS is typically time-series data collected by a variety of sensors, such as vibration signals, acoustic signals, currents and voltages. In addition, video, image, and text data are also used for equipment health monitoring, such as track defect detection [57] and equipment crack monitoring [58]. Currently, the PHM community has open-sourced dozens of datasets, such as bearing failure datasets [59], aircraft engine degradation datasets [60], three-phase motor failure datasets [61], industrial production monitoring datasets [62], and wind turbine monitoring dataset [63]. As a classical fault diagnosis dataset, the CWRU bearing datasets [64] contain only a limited number of operating conditions and four different health levels, each level encompassing only three failure levels. Depending on the operating conditions, the MFPT datasets [65] collect bearing data for three health states, as well as ten outer ring faults and seven inner ring faults at different loads. The DIRG datasets [66] are designed specifically for testing high-speed aerospace bearings and contains three health states: inner ring fault, rolling element fault, and health states. The C-MAPSS datasets [67] comprise operational data from four different types of aircraft engines,

consisting of 21 sensor signals used to predict the RUL of the engines. This dataset includes four subsets, each with varying numbers of operating conditions and fault scenarios.

Obviously, these datasets are small in the scale, which makes it difficult to meet the training and optimization requirements of LFMs. The emergence of the industrial Internet and the IoT has led to the installation of numerous sensors on modern industrial production equipment and various complex mechanical devices, which enables real-time monitoring of various physical quantities of the system and timely detection of abnormal conditions. As a result, most enterprises have collected a large amount of industrial data and established data centers. This industrial data may include various data such as sensor signals, images, and videos, as well as a large amount of textual information such as maintenance work orders and reports. Therefore, building LFMs to effectively utilize these multi-source heterogeneous data presents new challenges. In addition, these data are usually in the hands of equipment operators and may involve trade secrets, thus requiring the development of solutions that comply with stringent privacy protection regulations. Federated Learning [68], [69], [70] is a distributed machine learning framework with privacy-preserving and secure encryption features. It allows decentralized participants to collaborate in building machine learning models without disclosing private data, providing a viable approach to address data privacy and security issues.

B. Downstream Tasks of LFMs of ICPS

Fault diagnosis and RUL prediction form the backbone of a PHM system that can significantly improve maintenance practices, cost savings and asset management. These technologies not only ensure timely and accurate problem detection, but also provide valuable foresight for proactive decision-making and strategic planning. Therefore, robust fault diagnosis and RUL prediction technologies are essential to the safe operation and development of ICPS. We review deep learning-based methods for fault diagnosis and RUL prediction, as shown in Table I.

1) Fault Diagnosis for ICPS: Fault diagnosis is a key part of ensuring the safe and reliable operation of industrial equipment. With the development of IoT and AI technologies, deep learning-based fault diagnosis methods have gradually become a hotspot for research by its efficiency, accuracy and adaptivity. According to the data environment, fault diagnosis applications are categorized into four situations.

In ICPS fault diagnosis, vibration signals are usually converted to spectral images to obtain more comprehensive data features. Traditional conversion algorithms rely on a priori knowledge, such as continuous wavelet transform and Fourier transform. To fill this gap, Bai et al. proposed a spectral Markov transition field algorithm that does not require a priori knowledge of the parameters and is able to directly convert acceleration to spectral images and input them to the CNN [71]. Besides frequency domain transition, Kim et al. proposed a health-adaptive time-scale representation embedded CNN, utilizing a multi-scale convolutional filter to construct time-domain 2D input signals, which enhances the feature extraction [72]. To enhance the classification ability of the model, Xu et al. developed a hybrid deep learning model based on CNN and gcForest [73]. This method uses CNN to extract fault features, and the forest classification layer generates multiple decision trees to diagnose faults based on the features of each sub-dataset. The weak nature of incipient faults is a major challenge for fault diagnosis. Gao et al. proposed to address the weak nature of incipient faults by using an autoencoder to extract a feature representation of amplitude and phase information, and then augmenting the feature representation with useful structural information by capturing internal correlations [85].

In practice, machines operate normally most of the time, with faults occurring only occasionally, so healthy data far outweighs fault data. To address the data imbalance problem, Liu et al. proposed a deep feature generation network fault diagnosis method [79], which uses an attention mechanism to generate features that supplement fault data. To prevent the generated samples from being too similar to real samples and causing model collapse, a pull-away function is integrated to design a new objective function of the generator, ensuring the diversity of the generated data. When only health data is available, the model is unable to generate new data based on the existing data. Pan et al. proposed a method to generate pseudo-data features by combining fault indices [80]. The monitoring data and fault indices of a machine change with the machine state, and this property is utilized to construct 16-D time-domain fault indices (i.e., kurtosis, variance, skewness, etc.) and generate data on common fault types by modifying the time-domain fault indices. In the case of zero fault data, Hu et al. proposes a fault diagnosis model based on siamese convolutional autoencoder [82]. The corresponding negative samples are first constructed for the positive samples and then fed into the corresponding feature extraction networks respectively, so that positive and negative samples are far away from each other in the representation space, thus avoiding the negative effect caused by the small samples.

Noise in ICPS is unavoidable and is typically caused by various factors, such as sensor noise due to system errors or environmental conditions, data noise from electromagnetic interference or signal attenuation, and data loss and incompleteness. Noise negatively impacts the accuracy and reliability of diagnostics. To enhance the noise resistance of models, Su et al. proposed a method to artificially create noise data [83]. This method involves adding noise vectors to an autoencoder and using maximum mean difference as the loss function to reconstruct the original data. These reconstructed data are then used to train the noise resistance capabilities of the CNN. In addition to environmental noise, there are label noise in the data. To address this issue, Zhang et al. proposed a deep residual network based on an adaptive loss-weighted meta-network [110]. This network consists of a classification network and a meta-network. The meta-network is cross-trained with clean and noise labels and records the gradients from training with clean labels. The classification network is trained using noise labels and updates its parameters jointly based on the gradients recorded by the meta-network to build resistance to label noise.

Most existing fault diagnosis methods use non-Euclidean structured data, focusing on the correlation between adjacent sampling points, but they overlook the interactions between components and equipment in ICPS. The introduction of graph networks addresses this shortcoming [111]. Recently, an increasing number of studies have transformed industrial data into

TABLE I
SEVEN DEEP LEARNING FRAMEWORKS FOR PHM IN ICPS

Model	Related Work	Advantage	Disadvantage
Convolution Neural Network	[71], [72], [73] [74], [75]	(1) Adaptive feature extraction (2) High portability (3) Rich improvement strategies	(1) Limited context capture capability (2) Lack of interpretability (3) Requires large amounts of labelled data
Recurrent Neural Network	[76], [77], [78]	(1) Temporal correlation model (2) Adaptation to dynamic input length	(1) Difficult to train and implement (2) Gradient vanishing and gradient explosion (3) Parallel computing difficulties
Generative Adversarial Network	[79], [80], [81]	(1) High-quality generation of samples (2) No need for explicit probabilistic models (3) Wide range of application areas	(1) Training is unstable and difficult to tune (2) High computational resource requirements
Auto Encoder	[82], [83], [84] [85]	(1) Non-dependent data labels (2) Multi-optimization variants for various problems (3) Easy to implement	(1) Needs a pre-training process (2) Insufficient capture of information relevance
Graph Neural Network	[86], [87], [88] [89], [90] [91]	(1) Highly interpretable (2) Ability to reasoning	(1) The size of graph is arbitrary (2) The topology of a graph is complex
Knowledge Graph	[92], [93], [94] [95], [96], [97] [98], [99], [100]	(1) Structured information representation (2) Good semantic understanding (3) Data integration and high scalability	(1) Difficulty in dynamic updating (2) High computational complexity (3) Complexity of the build process
Transformer	[101], [102], [103] [104], [105], [106] [107], [108], [109]	(1) Global attention mechanism (2) Parallel computing and good scalability (3) Capturing long-distance dependencies	(1) Large number of parameters (2) Sensitive to input sequence length (3) High computational complexity

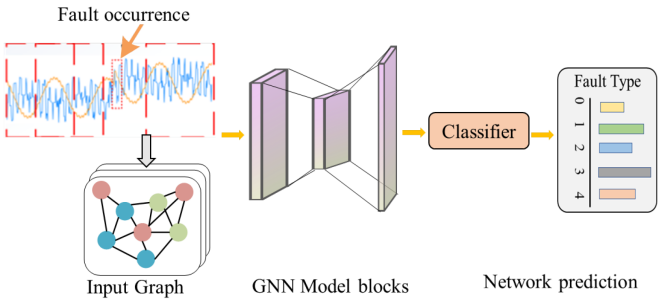

Fig. 5. Fault diagnosis framework based on GNN.

graph structures, considering the interactions between signals, and utilizing graph neural networks (GNN) to model fault patterns [112]. As shown in Fig. 5, GNN-based fault diagnosis is a node or graph classification task [86].

Chen et al. proposed an interaction-aware graph neural network for fault diagnosis [87]. An interaction-aware module was designed for exploring the interrelationships among industrial components, showing excellent reliability and robustness. In addition to edge synthesis based on the similarity between data points, Li et al. introduced a horizontally visual synthesis-based fault diagnosis model [88]. For each data sample, this model considers the data size as height and connects data nodes when there are no higher nodes obstructing between them. To address the issue of data noise interference, Wang et al. proposed a causal-trivial attention graph neural network fault diagnosis method [89]. This method learns causal subgraphs through causal decoupling to mitigate the confounding effects caused by noise, thereby improving the stability of the model. To enhance the generalization performance of fault diagnosis models in ICPS, Zhao et al. proposed a new semi-supervised GNN that integrates labeled and unlabeled information of the equipment [90]. This model first transforms 1-D data into spectral signals using fast fourier transform, then constructs a graph network from vector similarity-linked nodes, and finally feeds the sample data graph into graph convolutional layers and pooling layers for feature extraction and classification.

2) RUL Prediction for ICPS: Utilizing the historical data of the equipment to check its status can be more agile and efficient to detect the abnormal problems that may occur in operation. The deep learning-based RUL prediction method has the advantages of high accuracy, automation, robustness, which can significantly improve the prediction and health management level of industrial systems, and has a broad application prospect in ICPS. The core elements of RUL prediction include the construction of health indicators (HI) and the analysis of degradation trends. In terms of HI construction, Peng et al. established a DBN-based system fault feature representation method, taking the distance between the degraded state and the failure state as HI [113]. Chen et al. proposed a deep convolutional self-encoder model for adaptive construction of HI of rolling bearings [84]. Huang et al. utilized a bi-directional LSTM network for adaptively extracting, rotating, and fusing the wear features in the raw monitoring signals to construct a computerized numerical control machine HI [76].

To fully capture the time-series information of sensor signals, Chen et al. first extracted time-frequency domain features from the multi-channel signals collected by sensors [77]. Then, a sliding window approach was used to extract wear data, which, along with the time-frequency domain features, was modeled into a LSTM network to adequately consider the time-series characteristics of the data. To mitigate the adverse effects of sudden bearing failures, Cao employed the empirical mode decomposition method to decompose the raw vibration signals and selected

components based on the kurtosis criterion to reconstruct the vibration signal [114]. After filtering out the noise, time-domain and frequency-domain features were extracted from the reconstructed vibration signals. Most data-driven RUL prediction methods are unable to distinguish the contribution of different sensor and time-step data, which reduces data utilization. In this context, Song et al. proposed a time-series convolutional network based on distributed attention [74]. The method is based on a distributed attention mechanism that weights different industrial sensors and time steps separately. Then, the temporal convolution module with shared weights is used for feature extraction of the time series.

To accurately estimate health status without identifying the mathematical model of the system, Qin et al. proposed a LSTM network with macro-micro attention mechanisms [78]. The method begins by calculating typical features of vibration signals, such as mean, standard deviation, and kurtosis, etc. These features' principal components are then extracted using isometric mapping. By integrating these features, the method can effectively predict the health status of gear vibration signals. To fully utilize degradation information, Wen et al. proposed a hybrid RUL prediction method [75]. First, genetic programming is used to integrate physical sensor data into a composite HI, generating a clear nonlinear data-level fusion model. Next, using the belief function theory framework, RUL prediction is synthesized from each physical sensor and the developed HI as a decision-level fusion method, significantly reducing uncertainty. The success of RUL prediction relies on abundant operational failure data. However, in practice, such data may be insufficient. To address this issue, Zhang et al. employed a dual-channel fused convolutional recurrent neural network with a generative adversarial network to ensure high-quality data generation [81]. Considering the data privacy requirements and domain drift phenomenon of distributed multi-client collaborative training, Zhang et al. designed a multi-hop graph pooling adversarial network to reduce domain differences through adversarial transfer while achieving global modeling of input data [91]. Based on this, a distributed federated learning-based model consistency strategy was designed. This strategy dynamically allocates model weights to improve generalization ability while ensuring the privacy and security of each client local data.

Current data-driven PHM methods typically rely on large amounts of high-quality labeled data for training. However, in real industrial environments, the low incidence of equipment failures results in prominent issues of data scarcity and imbalance. Additionally, industrial data often contains noise and outliers, posing challenges to the robustness of models. The heterogeneity and complexity of the data further increase the difficulty of model training and application. Variations in data distribution across different equipment and operating conditions hinder the direct transferability of models to new environments or equipment.

Firstly, data augmentation techniques can be employed to generate more training data by transforming and expanding existing data, thereby alleviating data scarcity issues [26]. Generative adversarial networks can also be utilized to create high-quality synthetic data, further enriching the training dataset [115]. Secondly, active learning techniques can be adopted to interactively select the most valuable samples for labeling with the help of experts, reducing the dependence on large amounts of labeled data [116], [117]. In cases of data imbalance, oversampling and undersampling techniques can be used to balance the proportion of positive and negative samples in the dataset, preventing the model from being biased towards the majority class. Furthermore, Few-shot Learning and transfer learning techniques can be leveraged to train effective models even in scenarios with limited data [118]. Multisource heterogeneous data fusion is also a crucial approach to enhancing model performance. By incorporating real-time data collection and online learning techniques, models can be dynamically updated and adapted to new environments and equipment, thereby enhancing their multimodal learning capabilities.

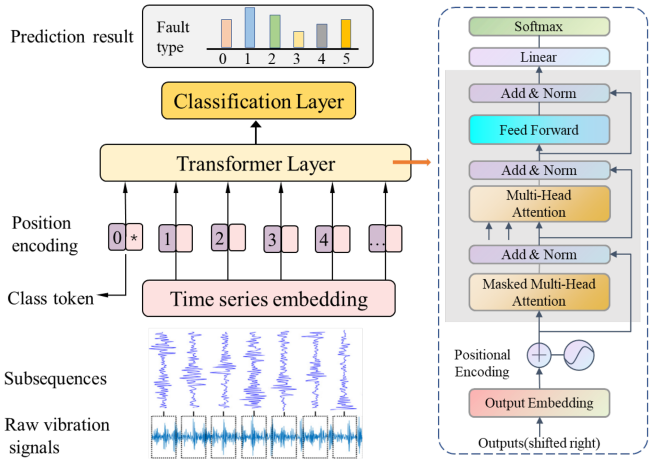

Fig. 6. Fault diagnosis framework based on Transformer.

C. Transformer for LFMs in ICPS

CNN and RNN are two of the most common architectures in ICPS, but each has its own weaknesses. Firstly, CNN is unable to capture relationships between targets, treat all pixels equally, and lack precise target localization capabilities [119]. Additionally, due to the local receptive field of convolutional kernels, a large number of convolutional layers must be stacked to obtain global information [120]. Secondly, RNN is not suitable for parallel computing, leading to inefficient training on large-scale datasets, and they still struggle to completely solve the problem of long-distance dependencies, making it difficult to establish effective connections over long sequences [121]. In contrast, Transformers excel in modeling long-distance dependencies, making them highly adapted to analyze and process various sensor data in ICPS. Research on fault diagnosis based on Transformers has now become a key research area, as illustrated in Fig. 6.

Compared with CNN and RNN models, Transformer-based fault diagnosis methods better extract temporal information from sensor signals, construct long-range dependencies, and significantly improve prediction accuracy [101], [102]. Wang

et al. proposed a Transformer-based high-speed train wheel wear prediction model [103], which effectively encodes both global and local information by leveraging the strengths of both Transformer and CNN. Fang et al. introduced an optimized lightweight Transformer framework that achieves efficient and accurate fault diagnosis while reducing computational complexity [104]. Li et al. developed a fault diagnosis method based on an improved attention mechanism, enabling the deep network to focus on information-rich data segments and ignore those that contribute less to the output [105].

Existing methods have difficulties in extracting long sequence temporal information from sensors, and Transformer-based RUL prediction methods are an effective solution. Li et al. proposed a convolutional dual-channel Transformer model with time window concatenation [106]. This model can mine long-term relationships and extract equipment degradation information from both the time and frequency domains without relying on any loop structures. Zhang et al. proposed ual-aspect self-attention based on transformer consisting of two encoders that can simultaneously extract features from different sensors and time steps [107]. The network is based only on the self-attention mechanism, processes long data sequences more efficiently, and adaptively learns to focus on the more important parts of the input. Considering the lack of ICPS fault data, Zhang et al. proposed a multilayer cross-domain transformer bearing RUL prediction method [108]. The method is able to capture simulated life cycle data through a dynamic model of the degradation process, while considering the loss of mutual information and retaining the generalized predictive knowledge of the measured data. Ding et al. proposed a Transformer-based multi-source domain generalization learning method [109]. The method can extract generalised degradation feature representations from multiple fault datasets under different known operating conditions or equipment conditions to assist the forecasting task in real application scenarios.

Although Transformer have achieved great success in ICPS, it still have some limitations that need to be addressed. Firstly, Transformer models are primarily designed for processing static input data such as text, and while they can incorporate temporal information through positional encoding, they do not directly consider time information. Therefore, when dealing with industrial time-series data, Transformers may struggle to fully learn the continuous temporal relationships in the data. Secondly, Transformer models may perform poorly when handling noisy industrial data. In real industrial production, noise in the data is inevitable, and Transformer models may not be sufficient to effectively manage this noisy data. Thirdly, industrial data typically includes various types of sensor data and a large amount of textual information, and Transformers generally find it challenging to simultaneously process extensive sensor data and heterogeneous data from multiple sources.

D. Knowledge-Enhanced LFMs for ICPS

A large amount of mechanistic, data and empirical knowledge has been accumulated in ICPS, which contributes to improve the interpretability and reliability of the model and supports subsequent model optimization. However, due to the different forms of presentation, this knowledge has not been adequately inherited and learnt in data-driven methods. To address this issue, scholars have proposed a series of knowledge-enhanced PHM studies for ICPS, such as fault dictionaries, fault trees, and Bayesian networks. The knowledge engineering oriented intelligent fault diagnosis system architecture is shown in Fig. 7.

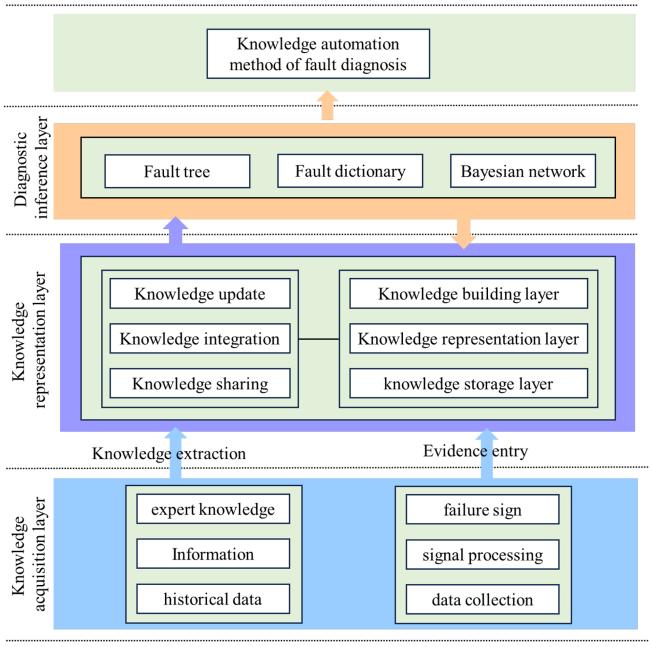

Fig. 7. The architecture of knowledge engineering oriented intelligent fault diagnosis.

Fault dictionary is a method of systematically summarizing information about equipment failure modes and characteristics, similar to dictionary form, presented in tabular form [122]. These fault dictionaries can be simple descriptive relationships between fault modes and fault features, complex nonlinear relationships, or even fuzzy relationships between equipment fault modes and their feature vectors. Fault dictionary-based diagnosis methods have the advantages of computational simplicity, well-defined relationships, and applicability to both linear and nonlinear systems, making them very suitable for PHM of ICPS equipment. Fault tree is a method that describes the logical relationship of events in a system through a causality tree diagram [123]. This method uses three elements: logic gates, input events, and output events to describe the structural relationships of complex equipment. Causality tree diagram can visually analyze the complex structure of the system and is very suitable for ICPS. Bayesian network analyzes the dependencies and correlation strengths between variables by means of joint probability distributions of the set of variables, and is suitable for expressing and reasoning about equipment uncertainty failure problems in ICPS [124].

Full-cycle PHM for ICPS includes activities such as design verification, manufacturing and testing, delivery and training, and operation and maintenance control. Managing and sharing vast amounts of information during the upstream and downstream phases of the lifecycle is a challenge [125]. Multiple

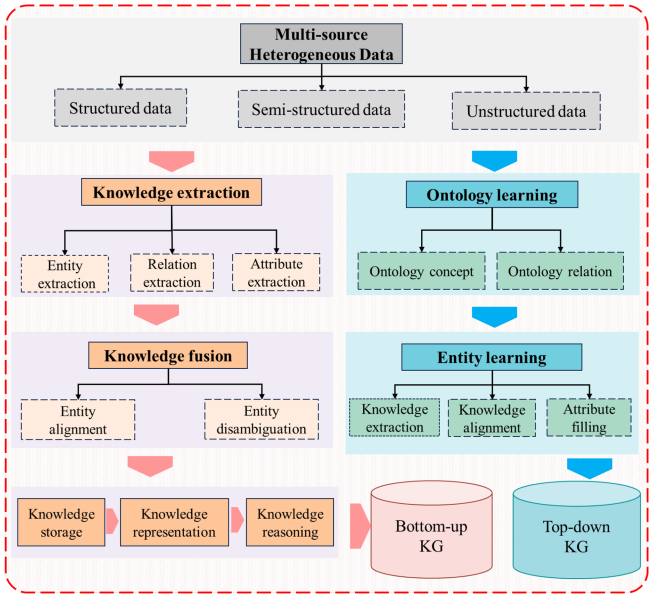

Fig. 8. Knowledge graph construction for ICPS.

sources of information and heterogeneous data pose significant challenges to the development of LFMs of PHM in ICPS. Heterogeneous data from various equipment and subsystems can be efficiently organized, stored and queried using knowledge graphs. This enables the construction of a unified data center that integrates information from multiple equipment, offering extensive application potential in ICPS scenarios such as aerospace, automotive, high-speed rail, and power systems [92].

Knowledge graph construction are generally categorized into top-down and bottom-up approaches [93], [94]. The top-down approach defines the framework of the domain knowledge system. The bottom-up approach enriches the knowledge content within the framework, which ensures the professionalism and accuracy of the knowledge system and covers the huge amount of data in the ICPS. Given the characteristics of knowledge and data in ICPS, the construction of knowledge graphs typically combines both approaches [95]. The construction of knowledge graph in ICPS includes ontology construction, knowledge extraction, knowledge fusion, and knowledge storage, and its process is shown in Fig. 8.

Knowledge graph-based PHM can be mainly divided into methods combined with Bayesian networks and methods combined with graph neural networks. The Bayesian network model has a network structure similar to the semantic web and the knowledge graph, and has the advantages of classification discrimination and quantitative analysis at the same time, which is an effective tool for assisting reasoning and diagnosis query [96], [97]. On the basis of constructing the graph structure through the relationship between entities in the knowledge graph, the graph neural network method is used to carry out fault cause analysis and fault localisation with significant advantages.

For fault analysis, Liu et al. proposed a new knowledge graph embedding and knowledge reasoning approach for fault diagnosis knowledge graphs [98]. The model explicitly learns an entity representation of the knowledge graph in an end-to-end learning mode, enabling automatic analysis of accident causes. For fault localisation, Liu et al. leveraged a priori knowledge to associate defects in bolt pairs in transmission lines with semantic objects [99]. This approach initialises defect nodes by features extracted from the joint region of the bolt parent pair in the image and initialises semantic object nodes by features obtained from the semantic object detection region in the image. Han et al. proposed a novel relational model-oriented steel production line equipment FD knowledge graph (SPLEFD-KG) based on global contextual information [100]. The method introduces a graph neural network to compute the embedding vectors of new entities outside the SPLEFD-KG in order to successively fill in the missing entities, thus information-rich and accurate fault-related knowledge.

Knowledge graphs hold tremendous potential in enhancing digital manufacturing, equipment, and project lifecycle management [126]. On one hand, a knowledge graph itself serves as a knowledge base, offering advantages such as visualization and ease of querying. On the other hand, knowledge graphs excel in expressing relational connections. By leveraging the semantic expression capabilities of knowledge graphs, it is possible to construct graph structures embodying physical or information coupling relationships, thus exploring data and information in unique ways. Combining industrial process knowledge with a data-driven methods, embedded in the model through rules or constraints, can further improve the interpretability and of the model. Rules from expert knowledge and historical experience can further enhance the credibility of decision making in LFMs.

VI. Opportunities and Challenges for LFMs in ICPS

LFMs have the potential to ensure the stable operation and sustainable development of ICPS by enhancing PHM. This survey provides a detailed overview of the overall advantages of LFMs in ICPS. We describe the critical challenges that must be addressed to ensure safe deployment, as LFMs of PHM will operate in particularly high-risk environments compared to the LFMs in other fields.

A. Paradigm Shifts With LFMs in ICPS

1) Generalizability: LFMs can learn and extract widely applicable features from vast amounts of data, ensuring robust performance across various application scenarios and working conditions. Traditional models often optimize for specific scenarios and lack the flexibility required in the diverse environments encountered in ICPS. In contrast, LFMs can perform joint learning on different types of data, establishing more universal prediction and analysis capabilities that apply to a wide range of conditions and equipment types. Additionally, LFMs with strong generalization capabilities can better handle emerging fault patterns and unknown operating states, crucial for enhancing system robustness and stability.

2) Multimodal Support: One of the critical advantages of LFMs is the ability to support multimodal data. In ICPS, data sources are diverse, including sensor data, image data, audio

data, and text data, each containing critical information. Traditional single-modal models often fail to fully utilize these heterogeneous data sources. In contrast, LFMs powered by AI techniques, can integrate multimodal data to achieve more comprehensive and accurate health management. For example, in monitoring equipment operation, sensor data can provide real-time status information, image data can assist in detecting visual changes, audio data can capture abnormal sounds during operation, and text data can log operational and maintenance records. By integrating these different data types, large-scale foundational models can develop a more holistic understanding of equipment operating conditions, enhancing the accuracy of fault detection and prediction. Multimodal support not only improves data utilization but also enhances the model's adaptability and application range, providing robust technical support for ICPS health management.

3) Reliability and Efficiency: LFMs accurately predict equipment failures by analyzing historical and real-time data, enabling proactive maintenance and preventing unplanned downtime, thereby significantly improving system reliability. Predictive maintenance based on LFMs can optimize maintenance schedules, reducing unnecessary inspections and maintenance efforts, saving both human and material resources. Additionally, LFMs can monitor and optimize key parameters in the production process in real-time, enhancing process stability and product quality. In terms of efficiency, LFMs can intelligently manage resource allocation by dynamically adjusting resources according to the actual operating conditions and production demands, maximizing resource utilization. For instance, in production line scheduling, the models can optimize scheduling plans based on equipment status, reducing production wait times and resource wastage.

B. Challenges of LFMs in ICPS

1) Data Quality and Security: The variability of ICPS scenarios reduces data reusability, while LFMs require a large amount of data for effective training, which poses a great challenge for data collection and preprocessing. ICPS generate diverse types of data, including sensor readings, operational logs, and environmental data. These data often contain noise, missing values, and outliers that can affect model training and prediction accuracy. Data cleaning and preprocessing are time-consuming and resource-intensive, and cannot be overlooked. Furthermore, ICPS data typically includes sensitive information and commercial secrets, so strict data privacy and security measures must be taken. Data must be encrypted during transmission and storage to prevent leaks and tampering. As cyber-attacks become increasingly sophisticated and frequent, ICPS must implement robust defense mechanisms to ensure data security. This involves not only technical safeguards but also strict management policies and emergency plans. Ensuring data quality and security is crucial for the reliability and stability of ICPS.

2) Model Interpretability and Trustworthiness: Although LFMs are excellent at handling complex data and predicting potential faults, their "black-box" nature makes the decision-making process difficult to interpret. Engineers and managers need to understand the basis for model predictions in order to make sound maintenance and operational decisions. The lack of interpretability not only reduces model acceptance, but also hampers the ability to quickly identify the root cause of prediction errors, affecting the accuracy and timeliness of decisions. Model predictions must be highly reliable, especially in ICPS applications involving safety and high risk. To enhance model trustworthiness, extensive testing and validation are necessary, along with the establishment of stringent evaluation standards. Moreover, LFMs need self-diagnosis and updating capabilities to adapt to changing operating conditions and emerging faults.

Extensive knowledge about equipment structure, fault mechanisms, and historical data has been accumulated in ICPS. This knowledge can be transformed into rules or constraints and embedded into LFMs to enhance the transparency of predictions. By leveraging historical data for model validation and calibration, the consistency between model predictions and actual conditions can be ensured, thereby improving model interpretability and reliability. Additionally, integrating expert knowledge and rules derived from historical experience can further enhance the decision-making credibility of LFMs.

3) Development Costs: Developing and deploying LFMs typically requires substantial computational resources, including high-performance computing platforms and GPU clusters, resulting in high hardware costs and energy consumption. The training and optimization processes are complex, requiring specialized technical personnel for debugging and maintenance, which increases labor costs. In addition, the process of data collection, cleaning and labelling is time-consuming and expensive, especially in ICPS where data must often be sourced from multiple heterogeneous origins and undergo complex preprocessing. To ensure the robustness and reliability of the models, extensive testing and validation are required, further escalating development costs. Maintaining and updating LFMs to adapt to changing operating conditions and emerging fault patterns is also a long-term, resource-intensive process. Although cloud computing and distributed computing can alleviate some of the resource pressures, the high development costs remain a major obstacle to the adoption of LFMs of PHM in ICPS. Effectively managing these costs and finding cost-efficient solutions are critical challenges that need to be addressed in this field.

VII. Conclusion

By collecting and analyzing data between connected machine clusters, Industry 4.0 offers unprecedented prospects for PHMP, but it also brings genuine challenges to the development and application of modern PHM. Existing PHM methods are mostly data-driven, facing issues such as insufficient model generalization capability, poor interpretability, and lack of trustworthiness. LFMs with cross-domain knowledge are equipped with powerful generalization and multitasking capabilities to meet the reliability and generalization requirements of ICPS. Therefore, this survey provides a comprehensive overview of the technical characteristics and current progress of LFMs. The literature review reveals a lack of research on LFMs of PHM in ICPS, and currently, there are no viable solutions to build

LFMs of PHM specifically for ICPS. This survey explores how to construct LFMs of PHM for ICPS from three key aspects: datasets, models, and algorithms. Finally, it attempts to examine the challenges associated with LFMs of PHM modeling from a broader perspective, aiming to offer valuable insights and guidance for future research in ICPS.

REFERENCES

[1] K. Zhang, Y. Shi, S. Karnouskos, T. Sauter, H. Fang, and A. W. Colombo, "Advancements in industrial cyber-physical systems: An overview and perspectives," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 716–729, Jan. 2023.

[2] J. Chae, S. Lee, J. Jang, S. Hong, and K.-J. Park, "A survey and perspective on industrial cyber-physical systems (ICPS): From ICPS to ai-augmented ICPS," *IEEE Trans. Ind. Cyber- Phys. Syst.*, vol. 1, pp. 257–272, 2023.

[3] D. G. Pivoto, L. F. de Almeida, R. da Rosa Righi, J. J. Rodrigues, A. B. Lugli, and A. M. Alberti, "Cyber-physical systems architectures for industrial Internet of Things applications in industry 4.0: A literature review," *J. Manuf. Syst.*, vol. 58, pp. 176–192, 2021.

[4] A. Abid, M. T. Khan, and J. Iqbal, "A review on fault detection and diagnosis techniques: Basics and beyond," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3639–3664, Jun. 2021.

[5] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mech. Syst. Signal Process.*, vol. 107, pp. 241–265, 2018.

[6] O. Fink, Q. Wang, M. Svensén, P. Dersin, W.-J. Lee, and M. Ducoffe, "Potential, challenges and future directions for deep learning in prognostics and health management applications," *Eng. Appl. Artif. Intell.*, vol. 92, 2020, Art. no. 103678.

[7] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[8] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[12] A. Kirillov et al., "Segment anything," in *2023 IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[13] K. Singhal et al., "Towards expert-level medical question answering with large language models," 2023, *arXiv:2305.09617*.

[14] W. Xiong et al., "AI-GOMS: Large ai-driven global ocean modeling system," 2023, *arXiv:2308.03152*.

[15] J. Huang et al., "ERNIE-GeoL: A geography-and-language pre-trained model and its applications in baidu maps," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 3029–3039.

[16] Y.-F. Li, H. Wang, and M. Sun, "ChatGPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps," *Rel. Eng. Syst. Saf.*, vol. 243, 2024, Art. no. 109850.

[17] N. Agrawal and R. Kumar, "Security perspective analysis of industrial cyber physical systems (I-CPS): A decade-wide survey," *ISA Trans.*, vol. 130, pp. 10–24, 2022.

[18] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proc. IEEE*, vol. 100, no. 1, pp. 210–224, Jan. 2012.

[19] G. Xiong et al., "Cyber-physical-social system in intelligent transportation," *IEEE/CAA J. Automatica Sinica*, vol. 2, no. 3, pp. 320–333, Jul. 2015.

[20] X. Jin, W. M. Haddad, and T. Hayakawa, "An adaptive control architecture for cyber-physical system security in the face of sensor and actuator attacks and exogenous stochastic disturbances," in *Proc. IEEE 56th Annu. Conf. Decis. Control*, 2017, pp. 1380–1385.

[21] J. L. Daan Ji, C. Wang, and H. Dong, "A review: Data driven-based fault diagnosis and rul prediction of petroleum machinery and equipment," *Syst. Sci. Control Eng.*, vol. 9, no. 1, pp. 724–747, 2021.

[22] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.

[23] R. Li, W. J. Verhagen, and R. Curran, "A systematic methodology for prognostic and health management system architecture definition," *Rel. Eng. Syst. Saf.*, vol. 193, 2020, Art. no. 106598.

[24] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, 2019.

[25] G. Niu, E. Liu, X. Wang, P. Ziehl, and B. Zhang, "Enhanced discriminate feature learning deep residual CNN for multitask bearing fault diagnosis with information fusion," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 762–770, Jan. 2023.

[26] T. Zhang et al., "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA Trans.*, vol. 119, pp. 152–171, 2022.

[27] Z.-H. Liu, B.-L. Lu, H.-L. Wei, L. Chen, X.-H. Li, and C.-T. Wang, "A stacked auto-encoder based partial adversarial domain adaptation model for intelligent fault diagnosis of rotating machines," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6798–6809, Oct. 2021.

[28] C. Ferreira and G. Gonçalves, "Remaining useful life prediction and challenges: A literature review on the use of machine learning methods," *J. Manuf. Syst.*, vol. 63, pp. 550–562, 2022.

[29] B. Yang, R. Liu, and E. Zio, "Remaining useful life prediction based on a double-convolutional neural network architecture," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9521–9530, Dec. 2019.

[30] M. Ma and Z. Mao, "Deep-convolution-based LSTM network for remaining useful life prediction," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1658–1667, Mar. 2021.

[31] A. Vaswani, N. Shazeer, and E. A. Parmar, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[32] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[33] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[34] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," in *Proc. Open AI*, 2018.

[35] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.

[36] J. Ye et al., "A comprehensive capability analysis of GPT-3 and GPT-3.5 series models," 2023, *arXiv:2303.10420*.

[37] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[38] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[39] H. Touvron et al., "LLaMA 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[40] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10078–10093.

[41] R. Deng, C. Cui, and E. A. Quan Liu, "Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging," in *Med. Imag. Deep Learn.*, 2023.

[42] D. Danny et al., "PaLM-E: An embodied multimodal language model," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 8469–8488.

[43] J. A. OpenAI et al., "GPT-4 technical report," 2024, *arXiv:2303.08774*.

[44] H. Xue and F. D. Salim, "PromptCast: A new prompt-based learning paradigm for time series forecasting," *IEEE Trans. Knowl. Data Eng.*, 2023, pp. 1–14.

[45] M. Jin et al., "Time-LLM: Time series forecasting by reprogramming large language models," in *Proc. Int. Conf. Learn. Representations*, 2024.

[46] C. Chang, W. Peng, and T.-F. Chen, "LLM4TS: Aligning pre-trained LLMS as data-efficient time-series forecasters," 2023, *arXiv:2308.08469*.

[47] C. Sun, Y. Li, H. Li, and linda Qiao, "Test: Text prototype aligned embedding to activate LLM's ability for time series," 2023, *arXiv:2308.08241*.

[48] D. Cao et al., "Tempo: Prompt-based generative pre-trained transformer for time series forecasting," 2023, *arXiv:2310.04948*.

[49] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 19622–19635.

[50] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "Timegpt-1," 2023, *arXiv:2310.03589*.

[51] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal data meets LLM - explainable financial time series forecasting," 2023, *arXiv:2306.11025*.

[52] Q. Xie, W. Han, Y. Lai, M. Peng, and J. Huang, "The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges," 2023, *arXiv:2304.05351*.

[53] T. B. Brown and E. A. Mann, "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[54] H. Xue, B. P. Voutharoja, and F. D. Salim, "Leveraging language foundation models for human mobility forecasting," in *Proc. 30th Int. Conf. Adv. Geographic Inf. Syst.*, 2022, pp. 1–9.

[55] Y. Chi, Y. Dong, Z. J. Wang, F. R. Yu, and V. C. M. Leung, "Knowledge-based fault diagnosis in industrial Internet of Things: A survey," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 12886–12900, Aug. 2022.

[56] Z. Liu, M. Xiao, H. Zhu, and J. Li, "Acquisition of missile fault diagnosis knowledge based on incomplete information of flow graph," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 632, no. 3, 2021, Art. no. 032055.

[57] A. K. Singh, A. Swarup, A. Agarwal, and D. Singh, "Vision based rail track extraction and monitoring through drone imagery," *ICT Exp.*, vol. 5, no. 4, pp. 250–255, 2019.

[58] G. Bayar and T. Bilir, "A novel study for the estimation of crack propagation in concrete using machine learning algorithms," *Construction Building Mater.*, vol. 215, pp. 670–685, 2019.

[59] H. Huang and N. Baddour, "Bearing vibration data collected under time-varying rotational speed conditions," *Data Brief*, vol. 21, pp. 1745–1749, 2018.

[60] M. Arias Chao, C. Kulkarni, K. Goebel, and O. Fink, "Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics," *Data*, vol. 6, no. 1, 2021, Art. no. 5.

[61] W. Jung, S.-H. Yun, Y.-S. Lim, S. Cheong, and Y.-H. Park, "Vibration and current dataset of three-phase permanent magnet synchronous motors with stator faults," *Data Brief*, vol. 47, 2023, Art. no. 108952.

[62] J. Wang, C. Xu, Z. Yang, J. Zhang, and X. Li, "Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 4, pp. 587–596, Nov. 2020.

[63] D. Zappalá, N. Sarma, S. Djurović, C. Crabtree, A. Mohammad, and P. Tavner, "Electrical & mechanical diagnostic indicators of wind turbine induction generator rotor faults," *Renewable Energy*, vol. 131, pp. 14–24, 2019.

[64] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mech. Syst. Signal Process.*, vol. 64/65, pp. 100–131, 2015.

[65] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "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 Soc. Eur. Conf.*, 2016.

[66] A. P. Daga, A. Fasana, S. Marchesiello, and L. Garibaldi, "The politecnico di torino rolling bearing test rig: Description and analysis of open access data," *Mech. Syst. Signal Process.*, vol. 120, pp. 252–273, 2019.

[67] S. Vollert and A. Theissler, "Challenges of machine learning-based RUL prognosis: A review on nasa's c-mapss data set," in *Proc. 26th IEEE Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, 2021, pp. 1–8.

[68] A. Arunan, Y. Qin, X. Li, and C. Yuen, "A federated learning-based industrial health prognostics for heterogeneous edge devices using matched feature extraction," *IEEE Trans. Automat. Sci. Eng.*, 2023, pp. 1–15.

[69] B. Li and C. Zhao, "Federated zero-shot industrial fault diagnosis with cloud-shared semantic knowledge base," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11619–11630, Jul. 2023.

[70] Y. Li, Y. Chen, K. Zhu, C. Bai, and J. Zhang, "An effective federated learning verification strategy and its applications for fault diagnosis in industrial IoT systems," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 16835–16849, Sep. 2022.

[71] Y. Bai, J. Yang, J. Wang, Y. Zhao, and Q. Li, "Image representation of vibration signals and its application in intelligent compound fault diagnosis in railway vehicle wheelset-axlebox assemblies," *Mech. Syst. Signal Process.*, vol. 152, 2021, Art. no. 107421.

[72] Y. Kim, K. Na, and B. D. Youn, "A health-adaptive time-scale representation (HTSR) embedded convolutional neural network for gearbox fault diagnostics," *Mech. Syst. Signal Process.*, vol. 167, 2022, Art. no. 108575.

[73] Y. Xu, Z. Li, S. Wang, W. Li, T. Sarkodie-Gyan, and S. Feng, "A hybrid deep-learning model for fault diagnosis of rolling bearings," *Measurement*, vol. 169, 2021, Art. no. 108502.

[74] Y. Song, S. Gao, Y. Li, L. Jia, Q. Li, and F. Pang, "Distributed attention-based temporal convolutional network for remaining useful life prediction," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9594–9602, Jun. 2021.

[75] P. Wen, Y. Li, S. Chen, and S. Zhao, "Remaining useful life prediction of IIoT-enabled complex industrial systems with hybrid fusion of multiple information sources," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9045–9058, Jun. 2021.

[76] C.-G. Huang, X. Yin, H.-Z. Huang, and Y.-F. Li, "An enhanced deep learning-based fusion prognostic method for RUL prediction," *IEEE Trans. Rel.*, vol. 69, no. 3, pp. 1097–1109, Sep. 2020.

[77] C. Tong, Q. Zhu, Y. Feng, and Y. Wang, "Sliding window-based real-time remaining useful life prediction for milling tool," in *Proc. IEEE 6th Int. Conf. Ind. Cyber- Phys. Syst.*, 2023, pp. 1–5.

[78] Y. Qin, S. Xiang, Y. Chai, and H. Chen, "Macroscopic–microscopic attention in LSTM networks based on fusion features for gear remaining life prediction," *IEEE Trans. Ind. Electron.*, vol. 67, no. 12, pp. 10865–10875, Dec. 2020.

[79] S. Liu, H. Jiang, Z. Wu, and X. Li, "Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis," *Mech. Syst. Signal Process.*, vol. 163, 2022, Art. no. 108139.

[80] T. Pan, J. Chen, J. Xie, Z. Zhou, and S. He, "Deep feature generating network: A new method for intelligent fault detection of mechanical systems under class imbalance," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 6282–6293, Sep. 2021.

[81] X. Zhang, Y. Qin, C. Yuen, L. Jayasinghe, and X. Liu, "Time-series regeneration with convolutional recurrent generative adversarial network for remaining useful life estimation," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6820–6831, Oct. 2021.

[82] H.-X. Hu, C. Cao, Q. Hu, Y. Zhang, and Z.-Z. Lin, "A real-time bearing fault diagnosis model based on siamese convolutional autoencoder in industrial Internet of Things," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 3820–3831, Feb. 2024.

[83] H. Su, L. Xiang, A. Hu, Y. Xu, and X. Yang, "A novel method based on meta-learning for bearing fault diagnosis with small sample learning under different working conditions," *Mech. Syst. Signal Process.*, vol. 169, 2022, Art. no. 108765.

[84] D. Chen, Y. Qin, Y. Wang, and J. Zhou, "Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction," *ISA Trans.*, vol. 114, pp. 44–56, 2021.

[85] T. Gao, J. Yang, S. Jiang, and Y. Li, "An incipient fault diagnosis method based on complex convolutional self-attention autoencoder for analog circuits," *IEEE Trans. Ind. Electron.*, vol. 71, no. 8, pp. 9727–9736, Aug. 2024.

[86] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mech. Syst. Signal Process.*, vol. 168, 2022, Art. no. 108653.

[87] D. Chen, R. Liu, Q. Hu, and S. X. Ding, "Interaction-aware graph neural networks for fault diagnosis of complex industrial processes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6015–6028, Sep. 2023.

[88] C. Li, L. Mo, and R. Yan, "Rolling bearing fault diagnosis based on horizontal visibility graph and graph neural networks," in *Proc. Int. Conf. Sens., Meas. Data Anal. ERA Artif. Intell.*, 2020, pp. 275–279.

[89] H. Wang, R. Liu, S. X. Ding, Q. Hu, Z. Li, and H. Zhou, "Causal-trivial attention graph neural network for fault diagnosis of complex industrial processes," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 1987–1996, Feb. 2024.

[90] X. Zhao, M. Jia, and Z. Liu, "Semisupervised graph convolution deep belief network for fault diagnosis of electormechanical system with limited labeled data," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5450–5460, Aug. 2021.

[91] J. Zhang, J. Tian, P. Yan, S. Wu, H. Luo, and S. Yin, "Multi-hop graph pooling adversarial network for cross-domain remaining useful life prediction: A distributed federated learning perspective," *Rel. Eng. Syst. Saf.*, vol. 244, 2024, Art. no. 109950.

[92] J. Wang, X. Wang, C. Ma, and L. Kou, "A survey on the development status and application prospects of knowledge graph in smart grids," *IET Gener. Transmiss. Distrib.*, vol. 15, no. 3, pp. 383–407, 2021.

[93] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[94] X. Zhu et al., "Multi-modal knowledge graph construction and application: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 2, pp. 715–735, Feb. 2024.
[95] D. Huang, C. Zhu, X. Liu, B. Zhou, S. Luo, and W. Zheng, "Application of intelligent knowledge graph with emergency fault diagnosis for power maintenance," in *Proc. Int. Conf. Distrib. Comput. Optim. Techn. (ICDCOT)*, 2024, pp. 1–5.
[96] G. Yang and X. Gu, "Fault diagnosis of complex chemical processes based on enhanced naive Bayesian method," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4649–4658, Jul. 2020.
[97] C. Lou et al., "Research on diagnostic reasoning of cloud data center based on Bayesian network and knowledge graph," in *Proc. Prognostics Health Manage. Conf.*, 2022, pp. 283–288.
[98] Y. Liu, Y. Ma, Z. Mao, B. Jiang, J. Liu, and J. Bian, "TD-GAT: Graph neural network for fault diagnosis knowledge graph," in *Proc. China Automat. Congr.*, 2021, pp. 1734–1739.
[99] L. Liangshuai et al., "Method and device for identifying missing pin bolts in transmission lines based on graph knowledge reasoning," *J. Intell. Syst.*, vol. 18, no. 2, 2023, Art. no. 9.
[100] H. Han, J. Wang, and S. Chen, "Construction and evolution of fault diagnosis knowledge graph in industrial process," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 3522212.
[101] Y. Jin, L. Hou, and Y. Chen, "A time series transformer based method for the rotating machinery fault diagnosis," *Neurocomputing*, vol. 494, pp. 379–395, 2022.
[102] Y. Ding, M. Jia, Q. Miao, and Y. Cao, "A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings," *Mech. Syst. Signal Process.*, vol. 168, 2022, Art. no. 108616.
[103] H. Wang, T. Men, and Y.-F. Li, "Transformer for high-speed train wheel wear prediction with multiplex local-global temporal fusion," in *Proc. IEEE 21st Int. Conf. Softw. Qual. Rel. Secur. Companion*, 2021, pp. 1175–1176.
[104] H. Fang et al., "You can get smaller: A lightweight self-activation convolution unit modified by transformer for fault diagnosis," *Adv. Eng. Inform.*, vol. 55, 2023, Art. no. 101890.
[105] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, 2019.
[106] L. Jiang, T. Zhang, W. Lei, K. Zhuang, and Y. Li, "A new convolutional dual-channel transformer network with time window concatenation for remaining useful life prediction of rolling bearings," *Adv. Eng. Inform.*, vol. 56, 2023, Art. no. 101966.
[107] Z. Zhang, W. Song, and Q. Li, "Dual-aspect self-attention based on transformer for remaining useful life prediction," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2505711.
[108] Y. Zhang, K. Feng, J. C. Ji, K. Yu, Z. Ren, and Z. Liu, "Dynamic model-assisted bearing remaining useful life prediction using the cross-domain transformer network," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 2, pp. 1070–1080, Apr. 2023.
[109] N. Ding, H. Li, Q. Xin, B. Wu, and D. Jiang, "Multi-source domain generalization for degradation monitoring of journal bearings under unseen conditions," *Rel. Eng. Syst. Saf.*, vol. 230, 2023, Art. no. 108966.
[110] K. Zhang, B. Tang, L. Deng, Q. Tan, and H. Yu, "A fault diagnosis method for wind turbines gearbox based on adaptive loss weighted meta-resnet under noisy labels," *Mech. Syst. Signal Process.*, vol. 161, 2021, Art. no. 107963.
[111] K. Feng et al., "Digital twin enabled domain adversarial graph networks for bearing fault diagnosis," *IEEE Trans. Ind. Cyber- Phys. Syst.*, vol. 1, pp. 113–122, 2023.
[112] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *" IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[113] K. Peng, R. Jiao, J. Dong, and Y. Pi, "A deep belief network based health indicator construction and remaining useful life prediction using improved particle filter," *Neurocomputing*, vol. 361, pp. 19–28, 2019.
[114] S. Cao and Y E. A. Jiang, "Bearing remaining useful life prediction based on optimized support vector regression model with denoising technique," in *Proc. IEEE 4th Int. Conf. Ind. Cyber- Phys. Syst.*, 2021, pp. 667–672.
[115] W. Mao, Y. Liu, L. Ding, and Y. Li, "Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study," *IEEE Access*, vol. 7, pp. 9515–9530, 2019.
[116] S. Attestog, J. S. L. Senanayaka, H. Van Khang, and K. G. Robbersmyr, "Robust active learning multiple fault diagnosis of PMSM drives with sensorless control under dynamic operations and imbalanced datasets," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9291–9301, Sep. 2023.
[117] C. Fan, Q. Wu, Y. Zhao, and L. Mo, "Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance," *Appl. Energy*, vol. 356, 2024, Art. no. 122356.
[118] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Few-shot transfer learning for intelligent fault diagnosis of machine," *Measurement*, vol. 166, 2020, Art. no. 108202.
[119] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.
[120] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech*, pp. 5036–5040, 2020.
[121] J. F. Kolen and S. C. Kremer, *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*. Hoboken, NJ, USA: Wiley, 2001, pp. 237–243.
[122] Y. Kong, Z. Qin, T. Wang, M. Rao, Z. Feng, and F. Chu, "Data-driven dictionary design–based sparse classification method for intelligent fault diagnosis of planet bearings," *Struct. Health Monit.*, vol. 21, 2021, Art. no. 147592172110290.
[123] E. Ruijters and M. Stoelinga, "Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools," *Comput. Sci. Rev.*, vol. 15–16, pp. 29–62, 2015.
[124] H.-B. Jun and D. Kim, "A Bayesian network-based approach for fault analysis," *Expert Syst. Appl.*, vol. 81, pp. 332–348, 2017.
[125] K. Wang and A. Takahashi, "Semantic web based innovative design knowledge modeling for collaborative design," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5616–5624, 2012.
[126] Q. Chen, Q. Li, J. Wu, and C. Mao, "Application of knowledge graph in power system fault diagnosis and disposal: A critical review and perspectives," *Front. Energy Res.*, vol. 10, 2022.

Ruonan Liu (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2013, 2015, and 2019, respectively. She was a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 2019. She is currently an Associate Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. She is also an Alexander von Humboldt Fellow with the University of Duisburg-Essen, Duisburg, Germany. Her research interests include machine learning, intelligent manufacturing and intelligent unmanned system. She was the recipient of the 2021 Outstanding Paper Award by IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, recognized as one of the World's Top 2% Scientists by Stanford University consecutively from 2021 to now and selected in the Young Elite Scientist Sponsorship Program by CAST in 2022. She is an Associate Editor or Leading Guest Editor of IEEE TRANSACTIONS ON INDUSTRIAL CYBER-PHYSICAL SYSTEMS, and *Sustainable Energy Technologies and Assessments*.

Quanhu Zhang received the B.S. degree in computer science and technology in 2022 from Nanjing Forestry University, Nanjing, China, where he is currently working toward the M.S. degree in computer technology with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include deep learning, graph neural networks, and fault diagnosis.

Te Han (Member, IEEE) received the B.Sc. and Ph.D. degrees in energy and power engineering from Tsinghua University, Beijing, China, in 2015 and 2020, respectively. In 2019, he was a Visiting Scholar with the University of Alberta, Edmonton, AB, Canada. From 2020 to 2023, he was a Postdoctoral Research Fellow with Department of Industrial Engineering, Tsinghua University, Beijing. He is currently an Associate Professor with the School of Management, Beijing Institute of Technology, Beijing. He has authored or coauthored two books and more than 50 articles in technical journals and conference proceedings. 11 of his articles have been honored with the "ESI highly cited paper", and five articles has been honored with the "ESI hot paper" in the Web of Science. His research interests include scientific machine learning, deep learning, trustworthy AI, and applications to sustainable energy, industrial diagnostics, prognostics and health management. He has been recognized as one of the World's Top 2% Scientists by Stanford University consecutively from 2020 to 2022. He was the Guest Editor of internationally renowned journals, such as *Reliability Engineering & System Safety*, and *Journal of Risk and Reliability*. He is the Editor of *Applied Soft Computing*, and an Associate Editor for IEEE SENSORS JOURNAL. He is also an active peer reviewer for more than 50 prestigious journals. He has received numerous prestigious awards, including the Young Elite Scientists Sponsorship by the China Association for Science and Technology (CAST), Shuimu Scholar from Tsinghua University and Excellent Doctoral Thesis at Tsinghua University.

Boyuan Yang (Member, IEEE) received the B.A., M.A., and Ph.D. degrees in mechanical engineering from the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, in 2013, 2015, and 2019, respectively. He was a Research Associate with the School of Electrical and Electronic Engineering, University of Manchester, Manchester, U.K. In 2024, he joined the Center for Advanced Control and Smart Operations, Nanjing university, Nanjing, China. His research interests include intelligent manufacturing, machine learning, condition monitoring, and wind energy.

Weidong Zhang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control science and engineering from Zhejiang University, Hangzhou, China, in 1990, 1993, and 1996, respectively. From 2003 to 2004, he was an Alexander von Humboldt Fellow with the University of Stuttgart, Stuttgart, Germany. From 2013 to 2017, he was Deputy Dean with the Department of Automation. He was a Postdoctoral Fellow with Shanghai Jiaotong University, Shanghai China, where he joined as an Associate Professor in 1998, and has been a Full Professor, since 1999. He is currently the Director of the Engineering Research Center of Marine Automation Shanghai Municipal Education Commission, China. He has authored or coauthored more than 300 papers and one book, and has been recognized as an Elsevier most cited Researcher. His research interests include control theory, machine learning theory, and their applications in industry and autonomous systems. Dr. Zhang was the recipient of the National Science Fund for Distinguished Young Scholars, China.

Shen Yin (Fellow, IEEE) received the M.Sc. and Ph.D. (Dr.-Ing.) degrees from the University of Duisburg–Essen, Duisburg, Germany. He is currently the DNV Endowed Professor with the Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim, Norway. He was elected to the Norwegian Academy of Technological Sciences (NTVA). His research interests include safety, reliability of technical systems, system and control theory, data-driven and machine learning approaches, applications in cyber-physical systems, health diagnosis, medical technology, and sustainable energy.

Donghua Zhou (Fellow, IEEE) received the B.Eng., M.Sci., and Ph.D. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1985, 1988, and 1990, respectively. From 1995 to 1996, he was an Alexander von Humboldt Research Fellow with the University of Duisburg-Essen, Duisburg, Germany, and a Visiting Scholar with Yale University, New Haven, CT, USA, during 2001–2002. He was with Tsinghua University, Beijing, China, in 1996, where he was promoted to Full Professor in 1997 and was the Head of the Department of Automation, during 2008–2015. From 2015 to 2023, he was the Vice President of the Shandong University of Science and Technology, Qingdao, China, China. He has authored and coauthored more than 320 peer-reviewed international journal papers and nine monographs in the areas of fault diagnosis, fault-tolerant control, and operational safety evaluation. Dr. Zhou is an Associate Editor for *Journal of Process Control*, the Vice Chairman of the Chinese Association of Automation, and the TC Chair of the SAFEPROCESS Committee, CAA. He was also the NoC Chair of the 6th IFAC Symposium on SAFEPROCESS 2006. He is a Fellow of IET and CAA, a Member of IFAC TC on SAFEPROCESS.