IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received June 13, 2017, accepted July 17, 2017, date of publication July 25, 2017, date of current version September 19, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2731945

Industrial Big Data for Fault Diagnosis: Taxonomy, Review, and Applications

YAN XU[1], YANMING SUN[2], JIAFU WAN[3], XIAOLONG LIU[1], AND ZHITING SONG[1]
[1]School of Business Administration, South China University of Technology, Guangzhou 510641, China
[2]School of Business Administration, Guangzhou University, Guangzhou 510006, China
[3]Guangdong Provincial Key Laboratory of Precision Equipment and Manufacturing Technology, South China University of Technology, Guangzhou 510641, China

Corresponding author: Jiafu Wan (jiafuwan2017@163.com)

This work was supported in part by the Natural Science Foundation of China under Grant 71571072, in part by the Fundamental Research Funds for the Central Universities under Grant 2015ZZ079 and Grant x2jqD2170480, and in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2016A030313734 and Grant 2016A030313735.

ABSTRACT Fault diagnosis is an important topic both in practice and research. There is intense pressure on industrial systems to continue reducing unscheduled downtime, performance degradation, and safety hazards, which requires detecting and recovering from potential faults as early as possible. From the historical perspective, this paper divides fault diagnosis into previous research and industrial big data era. According to primary drivers, this paper classifies fault diagnosis into knowledge-driven, data-driven, and value-driven methods. Among them, the former two approaches belong to the previous research on fault diagnosis. They mainly depend on expert experience and shallow models to detect and extract failures from relatively small size data. With the continuous exponential growth of data, it is insufficient to mine valuable fault information from massive multi-source heterogeneous data. The huge diagnostic value embodied in industrial big data has driven the emergence of the third category, which belongs to fault diagnosis based on big data. It consists of big data processing and analysis corresponding to high efficiency, cost effectiveness, and generality, which can deal well with problems that previous methods faced. We introduce the concept of a device electrocardiogram from the perspective of applicability to outline the present status of fault diagnosis for big data, and compare it with traditional diagnostic system. We also discuss issues and challenges that need to be further considered. It would be great valuable to integrate or explore more advanced diagnostic methods to handle collected industrial big data and put them into practice to mine the huge hidden diagnostic value.

INDEX TERMS Fault diagnosis, industrial big data, value discovery, device electrocardiogram.

I. INTRODUCTION

In the context of Industry 4.0, related technologies, such as the Internet of Things [1], wireless sensor networks [2], and cloud computing [3]–[5], have developed rapidly. At the same time, data acquisition and storage become easier and easier, promoting the arrival of the era of industrial big data. In 2001, Gartner [6] summarized the characteristics of big data into '3V,' namely, Volume, Velocity and Variety. Further, IBM extended the connotation of the characteristics of big data by including Veracity and Value, making it '5V.' Correspondingly, it becomes focal for today's research to determine how to produce tremendous practical value quickly and accurately from massive data, which comes from multiple and heterogeneous sources at a rapid rate [7]–[10].

In modern industry, production equipment develops towards being extremely precise, efficient and intelligent. Small performance degradation or security risks may bring serious consequences. It is vitally important to have a valid diagnosis approach to ensure the safe operation of the equipment. Before the arrival of big data, the previous research on fault diagnosis mainly depended on the richness of domain knowledge, the accuracy of diagnostic models, and the completeness of data samples [11]. These methods have the advantages of simplicity, interpretability and ease of development. But they are susceptible to disturbances in the environment, and produce a tremendous computation pressure when facing large-scale complex systems.

With the exponential growth of monitoring data, fault diagnosis faces enormous challenges dealing with industrial big data. It is like an iceberg where only a small part of fault information floats on the surface. It is hard to use the previous diagnostic methods to explore the true hidden value [12], [13].

At this time, the problem of transforming the growing volumes of data into the value is a considerable issue. The problem mainly includes two aspects. The first one is how to diagnose and predict failures rapidly or even in real-time using novel processing systems. The second one is how to deeply dig out the 'big' value of big data by improving the existing methods or leveraging new ones.

There are excellent reviews of fault diagnosis. Some existing research divides it into model-based and historical data-based methods, and provides systematic analysis and comparison from both quantitative and qualitative perspectives [14]–[16]. Some also divides it into model-based, signal-based, knowledge-based, hybrid and active diagnosis methods [17], [18], with special attention on real-time diagnosis and fault tolerance. But there are no studies that take into account and sum up industrial big data for fault diagnosis. Accordingly, this paper integrates the fault diagnosis of big data with previous research, and classifies it into knowledge-driven, data-driven and value-driven methods, according to driving factors. The main contributions of this paper are three-fold as follows:

- We analyze and summarize recent advances in fault diagnosis, classify them in details and more comprehensively, and discuss the current problems and new challenges fault diagnosis faces in the big data era.
- We anatomize fault diagnosis based on industrial big data, named value-driven diagnosis according to the huge value hidden behind it, and compare it to previous diagnostic methods from different perspectives.
- We introduce a novel concept and an application of the Device electrocardiogram (DEKG), which is inspired by the electrocardiogram for humans and monitors the devices' condition through their heartbeats.

The rest of this paper is organized as follows. Section II proposes a taxonomy for fault diagnosis that takes industrial big data into consideration. Section III reviews previous studies on fault diagnosis from the perspective of big data. Section IV summarizes fault diagnosis based on industrial big data, and compares it with previous diagnosis methods from the horizontal and vertical perspective. In section V, a novel concept of the DEKG and its applications are introduced. Section VI discusses the difficulties and challenges of fault diagnosis in the big data era. Finally, the conclusion is drawn in Section VII.

II. CLASSIFICATION OF FAULT DIAGNOSIS

In this paper, fault diagnosis is divided into knowledge-driven, data-driven and value-driven methods, according to driving factors, as shown in Figure 1. Knowledge-driven fault diagnosis is based on mechanical principles or artificial experience, and focuses on causal relationships. It applies to situations having small amounts of inputs and outputs, where the processes of the system are easy to model, and experience is easy to collect [18]. Once the model is established, it can realize real-time diagnosis, but only for specific types of failures.

As the development of an industrial process shows a large size and complexity, data scales along with the increase in uncertain disturbances. It is hard to ensure the robustness of knowledge-driven diagnosis to disturbances, and the sensitivity to failures. And it is also difficult to distinguish between causes and effects. Moreover, both cost of labor and cost of computation get higher, but it is difficult to maintain diagnostic accuracy, which can even lower than before. Having these problems, increasing data drives the transformation of fault diagnostic methods.

Data-driven methods can effectively improve diagnostic accuracy and the degree of automation [19]. They do not rest on the richness of expert experience or the precision of the mechanism model. But they depend on the correctness and completeness of online and historical data. Data-driven methods obtain implicit information using signal processing or data mining, where mined features are used to characterize the health conditions of the monitored systems. They mainly involve correlation analysis.

With the further development of industrial equipment towards higher complexity, speeds and intelligence [20], [21], related advanced technologies [22]–[24] started to be applied extensively. The growth of monitoring data is currently explosive, so we are entering the industrial big data era. As professor Viktor [25] described in his book, the real value of big data, just like an iceberg, floats in the sea, but most of it is beneath the surface. The previous diagnosis methods mostly explain values above the surface, for example, information about occurred failures. And latent values can ensure the best operation state of equipment, and are mostly related to failures that may occur in a meantime, such as the possibility of failures, the severity of possible failures, and an optimal recovery strategy. Clearly, latent values are of much larger value than surface ones.

However, the previous research on fault diagnosis can hardly dig out latent values from complicated monitoring data. This leads us to explore an applicable way to break through the limits of data volumes and types, so that we can shift the focus from the surface of equipment failures to deeper issues. We regard such methods as value-driven diagnosis. In the following sections, we introduce the principles of each method and related recent research, comparatively analyze their diagnostic abilities, and highlight the practical use of industrial big data, and directions for future development.

III. PREVIOUS RESEARCH ON FAULT DIAGNOSIS

Fault diagnosis originated in the 1960s. Early research regarded signal processing techniques and statistical analysis as major tools, and preliminary used artificial intelligence to extract fault features. It is committed to building diagnosis widen applications, increase the anti-interference ability, and improve diagnosis precision. Previous fault diagnosis includes knowledge-driven and data-driven fault diagnosis [26], [27].

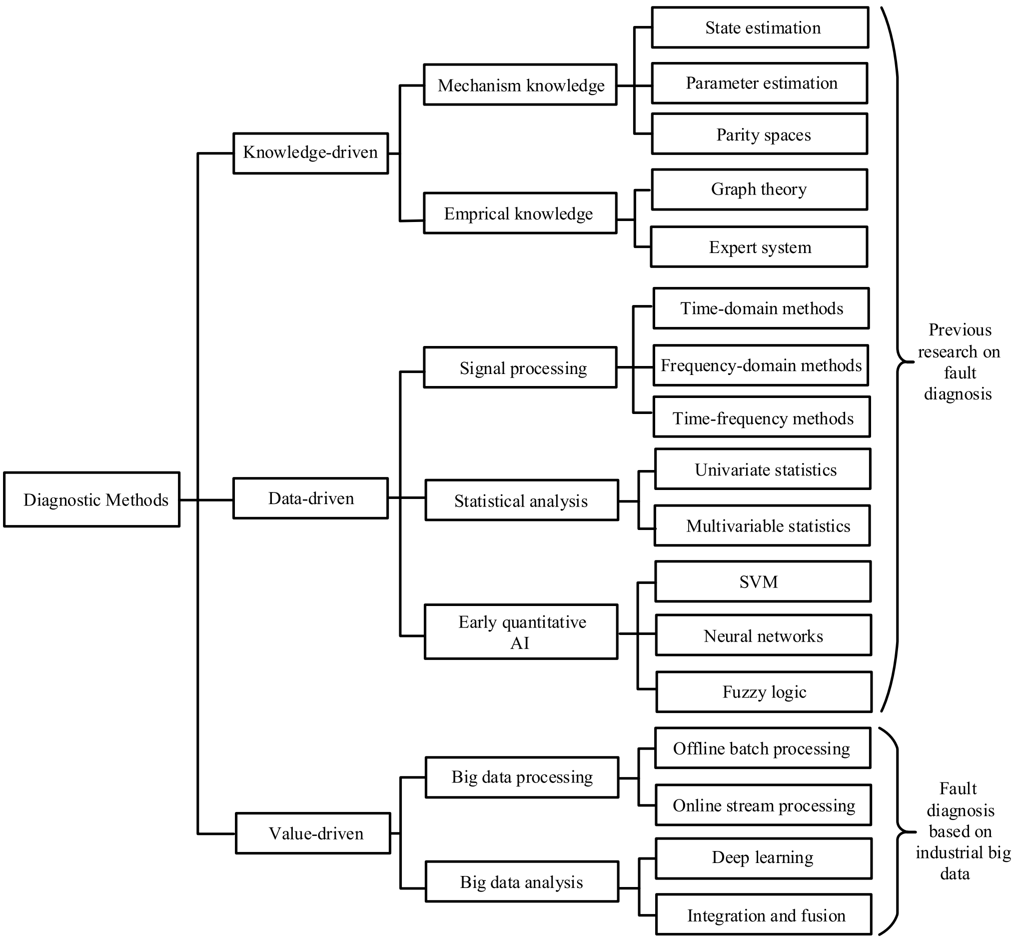

FIGURE 1. Classification of diagnostic methods.

A. KNOWLEDGE-DRIVEN FAULT DIAGNOSIS METHODS

Knowledge-driven fault diagnosis methods are built upon related knowledge as physical principles, fault mechanisms, and relevant expertise. They can identify the essence of the system, and realize real-time fault diagnosis. But the results of diagnosis are directly related to the precision of the mathematical model and the richness of experience. These methods can be further divided into mechanism knowledge-driven and empirical knowledge-driven methods according to different types of knowledge.

Mechanism knowledge-driven methods need to establish a precise mathematical model based upon the understanding of the physical mechanism and system structure. They construct a residual signal using the inputs and outputs of the system, in order to find inconsistencies between predicted and actual behaviors, so to detect, isolate and evaluate failures. They mainly include state estimation [28]–[31], parameter estimation [32], [33] and parity spaces [34], [35].

However, during a complicated dynamic industrial process, it is very difficult to build a mechanism model manually according to a deep insight into the system. It is always time-consuming, labor-intensive, and cannot maintain the validity of the model.

Empirical knowledge-driven fault diagnosis mainly depends on domain specific expertise and experience in long-term accumulation. It designs reasoning and decision-making mechanisms based on empirical knowledge to do diagnosis qualitatively, and can be further divided into the graph theory [36]–[38] and expert systems [39]–[41], according to different inference mechanisms.

This method can explain the key factors of a failure clearly and directly. It has less reliance on the mechanism model, and the results of diagnosis are easy to understand. But when a causal relationship is complex and variable, the inference process is extraordinarily complicated, so that it leads to massive computational time. Moreover, it leads to the problem of erroneous and missed diagnosis.

Therefore, the challenges knowledge-driven diagnosis faces in the big data era are mainly the following two aspects: (1) how to obtain comprehensive and reliable specialist knowledge to build a diagnostic model and inference mechanism, and (2) how to adaptively learn and diagnose new issues and compound problems that occur in the system.

TABLE 1. Improved methods of traditional fault diagnosis.

Author(s)	Method(s)	Diagnostic Objects	Advantages/Improvements
Ma and Yang [28]	High-gain observer Adaptive observer	MIMO non-linear uncertain systems	Without any linearization Independent of numerical optimization tools
Ma et al. [51]	Dwell-time based observers	Unknown inputs linear systems	Relaxed requirements for subsystems
Rundell et al. [52]	State and unknown input observers	Nonlinear systems with bounded exogenous inputs	Enhanced robustness Improved universalism
Zhai et al. [33]	Parameter estimation	MIMO closed-loop systems	Expanded range of applications Reduced complexity of models
Zhong et al. [34]	Parity space	Linear discrete time-varying systems	Reduced computation costs
Dong et al. [53]	Dynamic uncertain causality graph	Uncertain complex systems	Enhanced robustness Reduced computational complexity
Wang et al. [54]	Time Petri Nets	Timed discrete event systems	Reduced computational complexity
Zhao et al. [39]	Remote intelligent expert systems	Distributed systems	Increased precision Simplified procedure
Yan and Lu [44]	Improved Hilbert-Huang transform	Weak signals	Enhanced robustness Improved computational efficiency
Wang et al. [55]	Wavelet packet transform Manifold learning	Weak transient signals	Enhanced robustness Improved diagnostic performance
Yin et al. [46]	Improved PLS	KPI related processes	Increased precision Improved diagnostic efficiency
Shatnawi et al. [56]	Wavelet packet decomposition Extension neural network	Internal combustion engines	Enhanced expansibility Simplified models
Zhang et al. [57]	Incremental SVM	Roller bearings	Multi-fault diagnosis Indicated the severity
Talebi and Khorasani [47]	Neural network	Nonlinear systems	Enhanced robustness Indicated the severity

B. DATA-DRIVEN FAULT DIAGNOSIS METHODS

Data-driven fault diagnosis does not rest on building an explicit mathematical model based on prior knowledge, neither does it need to construct a reasoning mechanism based on expert experience. It uses different sorts of data mining techniques to extract and classify fault features in acquired vast operating data [42]. It mainly includes signal processing [43], [44], statistical analysis [45], [46] and early quantitative artificial intelligence methods [47].

Signal processing for diagnosis is aimed at extracting fault features in the time domain and frequency domain using various signal processing techniques. The operating features of a signal (e.g., a change in the amplitude or a phase drift) reflect the health status of the system to some extent [48]. Statistical analysis for diagnosis mainly uses statistics descriptions or statistical models to extract fault features. It mainly describes correlations among variables [49]. Quantitative artificial intelligence methods are used to train many types of learning algorithms using collected industrial process data, so that a computer can identify complex fault patterns automatically, and diagnose failures intelligently [50].

Above all, data-driven fault diagnosis can be appropriate to situations when the process mechanism is hard to master, the model and parameters are difficult to determine, and large amounts of data can be utilized. However, it requires high data-quality, and relies on the completeness and the representativeness of sample data. In addition, as the data volume and data types increase, the computational complexity rapidly increases. This produces expensive calculations. At the same time, this method can hardly handle real-time streaming data in a timely manner. And it also has limitations in the diagnosis of new and compound fault problems.

In allusion to the problems mentioned above, many improved methods are proposed in recent years. Table 1 shows improved methods and their performance in details. It can be seen that recent research involving previous methods gets inputs from specific equipment with a specific fault type or a particular system. And they mainly focus on the improvement of basic performance, including increasing precision, widening the scope, strengthening the robustness of uncertainty, and reducing the computational complexity.

But when facing massive, hierarchical and multidimensional information in the era of big data, it is necessary to further improve performance, such as the cost and speed of massive data processing, the generation ability of diagnosis, the depth of information mining, the integration of fault features, and the intelligence degree of fault diagnosis. Therefore, fault diagnosis more applicable to the characteristics of big data has become a hot topic of research.

IV. FAULT DIAGNOSIS BASED ON INDUSTRIAL BIG DATA

With the increasing degrees of automation and intelligence of industrial equipment, along with the development and widening of the application spectrum of related advanced technologies [58], [59], data started to grow exponentially. It becomes important to process and analyze collected industrial big data in order to obtain a great diagnostic value using it. However, corresponding problems, such as computational overload, widespread uncertainty, the lack of fault samples, and the difficulty of predicting fault types, have prevent the applications of previous fault diagnosis methods [60]. As a consequence, the question of how to realize rapid analysis and deeper mining of massive monitoring data has become a study focus in the big data era. In this section, we describe and summarize value-driven diagnosis in two ways: big data processing systems and big data analytical methods.

A. BIG DATA PROCESSING SYSTEMS

Traditional processing methods often rely on local high-performance computers and simple parallel operations to improve computational power. The key issues that need to be solved are overcoming the limitations of original relational databases and serial algorithms, and developing new types of distributed processing systems for big data [61].

A big data processing system with high efficiency and cost effectivity provides basic support for fault diagnosis based on industrial big data. At present, studies on big data processing systems are mainly focused on batch processing of historical monitoring data and real-time processing of online data. In what follows, we describe the characteristics and research advances of these two processing systems for fault diagnosis.

1) OFFLINE BATCH PROCESSING

The operation of large-scale complex industrial systems generates huge amount of monitoring data every day. As data volumes increase, requirements for software and hardware, and computational time becomes higher and higher [62]. How to uncover valuable information from accumulated massive data for fault prediction is a main concern. Offline batch processing is aimed at the problem mentioned above, and it is more suitable for mass stored data, as it focuses more on the accuracy and comprehensiveness of analysis rather than real-time diagnosis.

MapReduce [63] is a fairly typical big data batch processing model having the features of efficiency, cost control and scalability. It distributes mass data in a cluster of computers, where hardware resources used can be adjusted dynamically according to the analysis task. In addition, MapReduce can be used to handle unstructured data. It shields the low-level details of parallel processing, and, hence, reduces computational complexity and developer ability requirements.

All these features make MapReduce very popular in the field of fault diagnosis based on industrial big data. In [64], a novel framework named CloudView is proposed in a cloud computing environment. The jobs of case-base creation in this framework leverage the MapReduce parallel data processing model, where fault prediction can be done on a timescale of seconds. In [65], a MapReduce framework for fault diagnosis in cloud-based manufacturing is developed. It is used to recognize fault patterns automatically and effectively after solving a data imbalance problem. In [66], the MapReduce data processing mechanism is used to build a fault diagnosis model which can improve diagnostic efficiency and reduce fault costs in a photovoltaic power station.

However, MapReduce has the shortcoming of low computational efficiency in complex calculations due to iterative computations. In response to this issue, a distributed computing system called Spark [67] was developed in the University of California, Berkeley. It leverages memory to process data in order to reach faster processing. An integrated data preprocessing framework based on Spark is proposed in [68] for fault diagnosis. It can help to decrease data processing time while improve classification accuracy.

2) REAL-TIME STREAM PROCESSING

Batch processing has drawbacks in terms of process complexity and the lack of the ability to provide diagnosis feedback in real time. However, with time, the value contained in dynamic data flows decreases. Therefore, real-time monitoring is needed in variable and complex operating conditions in order to ensure timely maintenance. A stream processing system is an appropriate choice to realize this. Not only does it have fault tolerance and scalability, but it also can mine diagnostic value timely, and realize real-time diagnosis to keep equipment running constantly.

The Storm of Twitter [69] is currently often used as one of stream processing systems. In comparison with the batch processing architecture, Strom can handle infinite data flows quickly and reliably. It has the features of strong fault tolerance, powerful expansibility and low latency. And it is helpful to realize real-time monitoring, continuous diagnosis, performance improvement and remote maintenance.

Batch processing systems and streaming processes are often combined to provide comprehensive diagnosis online and offline. In [70], a manufacturing big data solution used for fault diagnosis is proposed. It includes a mechanism for real-time active maintenance (MRAM) based on Storm and a mechanism for off-line prediction and analysis (MOPA) based on Hadoop. In [71], a cloud-based data-intensive framework, which combines Storm and MapReduce, is proposed. It solves the problems of heterogeneous data and response times.

B. BIG DATA ANALYSIS METHODS

Under the basic support of a big data processing system, the implicated value of big data needs effective analytic methods to dig it out. Previous research on fault diagnosis generally uses relatively simple and shallow models to analyze and process single sample sets. These models have shown the lack of capability in performance and generalization. It is hard to satisfy the requirements of fault diagnosis based on big data. Accordingly, using deep learning to obtain the hierarchy representation of fault features, and using integrated methodology and data fusion techniques to obtain more reliable and comprehensive diagnostic information become a developing trend.

1) DEEP LEARNING

When extracting fault features from a large amount of monitoring data, the traditional artificial intelligence requires highly professional knowledge and a significant amount of engineering techniques to manually design classification.

And it also uses shallow models, which make it limited to gain an insight into raw data. To explore a general and automated feature extraction method, Hinton and Salakhutdinov [72] proposed deep learning, which is good at finding complex structures in high dimensional data. It can extract fault features adaptively using enough conversions and combinations, and distill the physical significance of features without manual intervention.

In [73], a multi-sensor fault diagnosis method based on deep belief learning (DBN) is proposed to provide effective health diagnosis. The DBN is composed of a stack of restricted Boltzmann machines. It takes unlabeled data as a training sample to train entire networks layer by layer, and then uses the BP algorithm to fine tune the parameters. Compared to the traditional processing of training models, the DBN has advantages in fast training, quick convergence, and the classification of unlabeled data. It solves the problem of extracting features from multi-sensor heterogeneous data with efficiency and accuracy.

The DBN is also used in a hierarchical diagnosis network (HDN) in [74]. The HDN can be divided into two functional layers, where the first layer is developed to identify and locate faults using the DBN, while the second layer also uses the DBN to rank fault severity. It expresses more details regarding diagnosis, and overcomes the overlapping problem caused by disturbances. In [75], intelligent fault diagnosis constructed for deep neural networks is proposed. It allows avoiding the dependence on traditional signal-processing techniques and artificial experience, achieves health diagnosis adaptively and automatically.

2) INTEGRATION AND FUSION

On the one hand, due to the complexity of equipment and the instability of environment, the occurrence of failures is always caused by the coupling of multiple factors. The traditional diagnostic methods cannot reflect the running state of equipment adequately, or even cause missed and erroneous diagnosis by analyzing one-dimensional data or from a single perspective. Therefore, it is necessary to fuse multi-source heterogeneous data for various aspects of diagnostic information mining. It can be divided into data-level, feature-level and decision-level fusion according to the progressive relation with the workflow of fault diagnosis.

In [76], fault diagnosis is regarded as a multi-sensor data fusion problem. It completes the process of fault signal classification by using the Support Vector Machine (SVM) and Short Term Fourier Transform (STFT) techniques. And with the increase in available sensor data, the diagnostic accuracy, reliability and robustness also increase. In [77], information fusion fault diagnosis using evidence reasoning is proposed. It resolves conflicts among evidence caused by uncertain factors, so that it reduces the influence of uncertain factors, and improves the accuracy of diagnosis results. In [78], the proposed fault diagnosis method breaks through the limitations of previous research that only provides diagnosis of individual components. It detects multiple faults by using combinational logic and fuzzy logic to reach systematic diagnosis.

On the other hand, there is no existing diagnostic method which would be able to detect the equipment health status and reach the demand of high confidence by its own. It is helpful to integrate knowledge-driven and data-driven methods in value-driven diagnosis. This way, it can strengthen the research on fault mechanisms with the support of data analysis, and promote the imagery cognition of faults to deeper reveal fault mechanisms, and, therefore, to interpret the mined diagnostic value in a more plausible way.

In [79], a novel distributed parallel computing model with the characteristics of both directed graphs and neurons is proposed. This model can visually express fuzzy rules in a fuzzy diagnosis knowledge base, and model dynamic reasoning processes timely. In [80], a combined self-diagnosis method is proposed to achieve higher automation and reliability. The Bayesian networks of it are used to establish a connection model, and the case-based reasoning of it is used to recognize fault features to reduce the complexity of the diagnostic process and human intervention. In [81], intelligent fault diagnosis combines data mining technologies and Bayesian networks. It takes advantages of data mining to avoid local optima in Bayesian networks, and also uses the merits of Bayesian networks to help data mining to handle incomplete and uncertain information.

The features of fault diagnosis based on industrial big data mentioned in this section are summarized in Table 2. The table shows that the objects of big data diagnosis are systems with massive data, and improvements in advanced performance mainly focus on the aspects of quick analysis, intelligence, automation and the deep diagnosis of hidden failures.

C. COMPARISON AND ANALYSIS OF PREVIOUS FAULT DIAGNOSIS

1) HORIZONTAL COMPARISON OF DIAGNOSTIC METHODS

All knowledge-driven, data-driven and value-driven diagnosis methods have their own advantages and disadvantages. Each method has specific situations that allow full exertion of advantages, but as circumstances change, the method may be not applicable. Figure 2 shows the phase of development of these methods and their applicable conditions.

When the data volume is small, knowledge-driven diagnosis performs best. It has the advantages of easy and straightforward inference, strong explanatory and rapid diagnosis. But data-driven methods as well as value-driven methods, cannot gain accurate fault features from small data sets.

As data becomes larger and richer, it is increasingly difficult to obtain expert experience and relevant knowledge, which hinders the applicability of knowledge-based diagnosis models [11]. However, data-driven methods are simple, and can provide diagnosis without modeling. This can help to exploit fault information accurately from large data sets and exclude the interference of uncertainty [16].

TABLE 2. Fault diagnosis on industrial big data.

Author(s)	Method(s)	Diagnostic Objects	Advantages/Improvements
Bahga and Madisetti [64]	CloudView framework	Massive machine maintenance data	Extensibility Prediction of faults on the timescale of seconds
Kumar et al. [65]	MapReduce framework	Cloud-based manufacturing	Ability to process highly imbalance unstructured data sets Automatic fault pattern recognition
Wan and Tang [70]	Storm Hadoop	Manufacturing big data	Efficient real-time maintenance Rapid prediction and analysis
Tamilselvan and Wang [73]	Deep belief network	Large complex systems	Fast inference Ability to encode richer and higher order network structures
Jia et al. [75]	Deep neural network	Rotating machinery with massive data	Lower dependence on human labor Ability to represent complex features
Banerjee and Das [76]	Multi-sensor data fusion technology	Multi-sensor systems	Improved diagnostic accuracy Insights into system behavior
Xiong et al. [77]	Information fusion technology	Large rotating machinery	Fusion of evidence conflicts Improved diagnostic accuracy
Peng et al. [79]	Parallel fuzzy reasoning SN P systems	Dynamic fuzzy systems	Visualizations Ability to handle fuzzy information
Bennacer et al. [80]	Bayesian networks Case-based reasoning	Virtual private networks	Reduced human intervention Increased level of automation

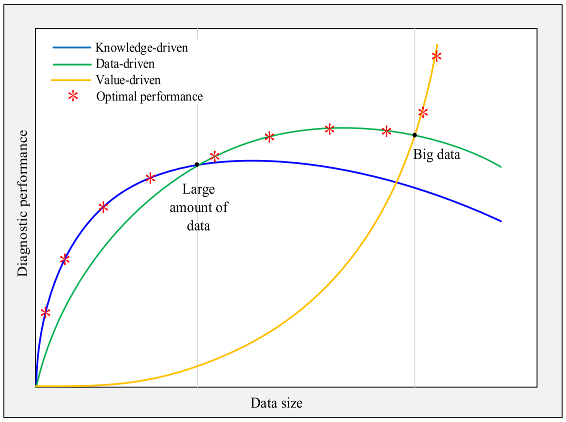

FIGURE 2. The phase of development and optimal performance of fault diagnosis methods.

Though value-driven methods have similarities with data-driven methods, both provide diagnosis by processing monitoring data. But using value-driven methods for this magnitude of data is a bit of overkill, as it makes the cost of computation too large and the initialization time too lengthy.

With further exponential increase in data, the previous fault diagnosis methods show problems, such as a too slow diagnostic speed, incomplete or inaccurate results, or even inability to produce a result. Value-driven diagnosis methods, which have excellent ability in the value mining of big data, show the greatest performance in this phase.

2) VERTICAL ANALYSIS OF DIAGNOSTIC PROCESSES

The process of fault diagnosis can be divided into three stages, namely, data acquisition, feature extraction and fault decision [17], from the perspective of the workflow of diagnosis. As shown in Table 3, every stage has its features and diagnostic tasks, but they are very different between previous methods and those for fault diagnosis based on big data.

In the stage of data acquisition, previous research mainly analyzes small scale and short time data collected by a few or even a single sensor. With the development of wireless sensor technologies, a variety of sensors perceive diverse data constantly. The resulting multivariate time series and incomplete or uncertain heterogeneous data are the first challenge fault diagnosis faces in the big data era [82]. Since the value of such data is high, we need to look for more effective and applicable methods to find out complex associations in it.

In the stage of feature extraction, previous fault diagnosis improves its computing abilities by changing computer with high performance or adding hardware (e.g., CPU and memory). This implies significant costs, though without useful diagnosis from data flows. Therefore, using a new class of processing systems, which are easily extensible and cost-effective, becomes a required trend in the era of industrial big data. Regarding the analytical approach in this stage, the demand for diagnosis has shifted from shallow analysis for a specific type of failures to deep and adaptive analysis for diverse and dynamic failures.

In the final phase of decision making, previous diagnostic methods output the results of diagnosis directly, due to the utilization of one-dimensional data and simple models. Diagnostic methods based on big data often use synthesized methods to process fused information for more comprehensive and well-founded decisions. Clearly, the results of big data analysis are worth more.

V. AN APPLICATION OF INDUSTRIAL BIG DATA FOR FAULT DIAGNOSIS

As is well known, a short time breakdown of industrial equipment can cause tremendous losses, especially in today's automotive, continuous and compact manufacturing. Traditional

TABLE 3. Comparison of previous fault diagnosis and diagnosis based on big data.

	Previous Fault Diagnosis	Fault Diagnosis Based on Big Data
DATA ACQUISITION	Single data source	Distributed data sources
	Structured data	Heterogeneous, multi-dimensional, uncertain, incomplete data
FEATURE EXTRACTION	Serial extraction based on prior knowledge	Adaptive parallel extraction
	Shallow structure	Deep structure
	Single model	Universal model
FAULT DECISION	Direct diagnosis	Integrated decision making based on information fusion
		Visualization of diagnostic results

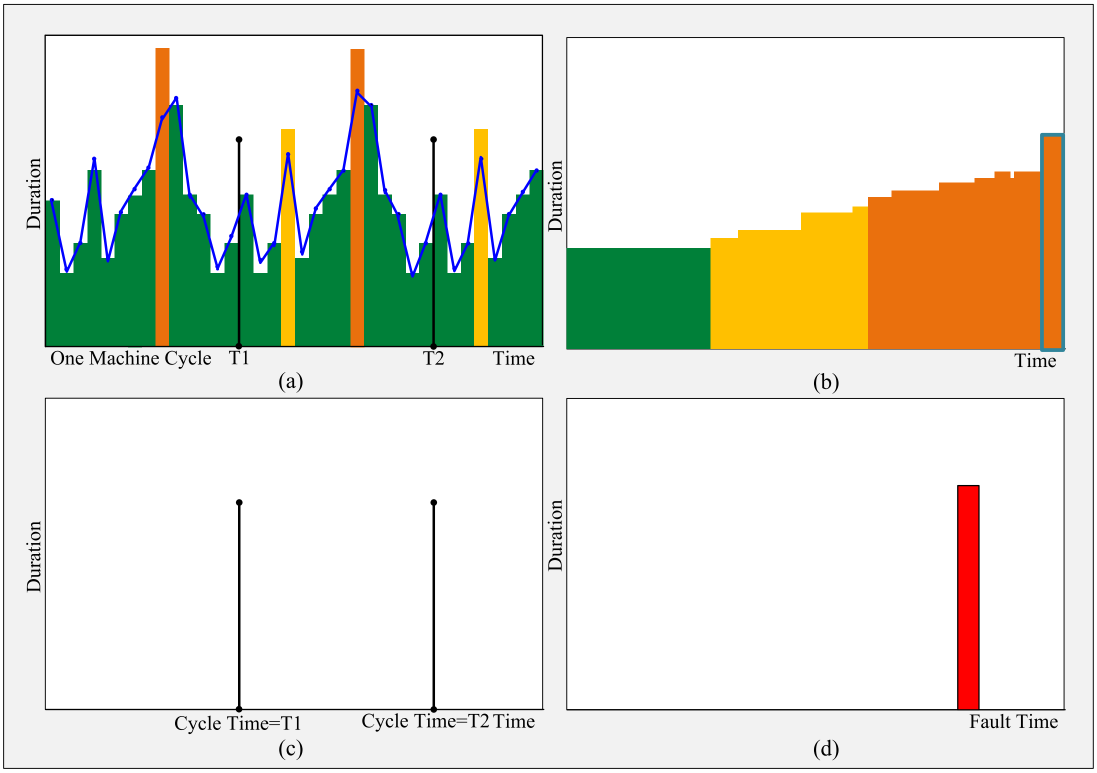

FIGURE 3. Comparison between the DEKG and traditional diagnostic processes. (a) DEKG of a machine. (b) DEKG of a movement. (c) Traditional monitoring chart. (d) Traditional alarm.

fault diagnosis often complies with the fail and fix (FAF) maintenance strategy, and it regards the failure of devices as a sudden shutdown problem. But in fact, there is a process of degradation before an unexpected stop. Therefore, to prevent failures that may occur at some time in the future, we can transform the maintenance strategy to predict and prevent (PAP) strategy by using advanced analytical and prediction technologies.

In recent years, some intelligent prognostic systems have been developed and implemented to predict the performance and life expectancy of machines, such as Watchdog Agent^{TM} [83] and SensorCloud^{TM} platforms [84]. Diffenet diagnostic system has its own features. For example, Watchdog Agent^{TM} could realize predictive condition-based maintenance through enabling multi-sensor assessment, and SensorCloud^{TM} could easily upload any data through their open data API in security. In what follows, we introduce a novel idea called ''Device electrocardiogram'' (DEKG) [85] to visualize the health conditions of monitored equipment.

A. DEVICE ELECTROCARDIOGRAM (DEKG)

Inspired by the electrocardiogram for humans, which presents invisible heartbeats as visible images, the DEKG aims at devices' heartbeats to visualize every event and motion of equipment. It can monitor the status of operations through the changes of DEKG, and predict downtimes, providing predictive and proactive maintenance.

Figure 3(a) shows the DEKG of a machine, where the blue baseline stands for the ideal conditions of equipment, and different color bars represent different health conditions. The bars under the baseline are green, meaning that the machine is running in a good status. They become yellow above the blue baseline within a threshold. This provides a reminder that the machine needs to be monitored. Beyond the threshold, bars turn orange. This warns that the device may break down in a meantime, and the problem must be solved as soon as possible. Moreover, the DEKG can depict the detailed performance of a certain motion or event, as shown in Figure 3(b). We can obtain the evolution of the warning component, and find the root causes of the problem.

For comparison, Figure 3(c) and Figure 3(d) show traditional monitoring charts. Figure 3(c) only shows the overall condition of a machine. We can only obtain the duration over a cycle, but have no understanding of detailed motions which may lead to failure. And a fault alarm is sent only after a failure occurs, as shown in Figure 3(d), which results in unrecoverable loss.

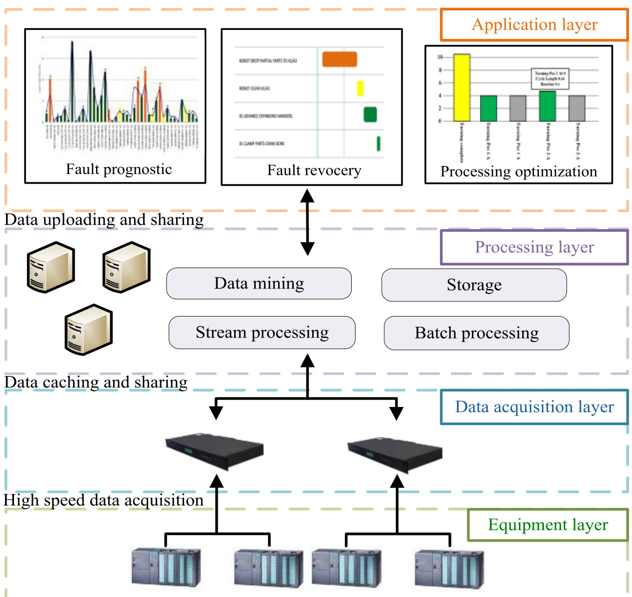

FIGURE 4. The architecture of the DEKG.

B. PRINCIPLES IN FAULT DIAGNOSIS

Figure 4 shows the architecture of a DEKG fault diagnosis system. It consist of four layers, namely, the equipment layer, data acquisition layer, processing layer and application layer. The data collectors in the data acquisition layer gather and format massive data using the PLC from the equipment layer, which can avoid requiring additional sensors. As the scanning periods of data collectors are the same as the PLC, the precise time of every move can be collected to capture valid changing details. Then, the formatted data is cached and shared to the processing layer. In this layer, some fault diagnostic methods are chosen to perform degradation assessment and prediction. When devices operate normally, it can also optimize the machine cycle to improve the throughput. Finally, the results of the analysis are uploaded to the application layer and produce the DEKG to help to predict, recover warning parts, and optimize the cycle time.

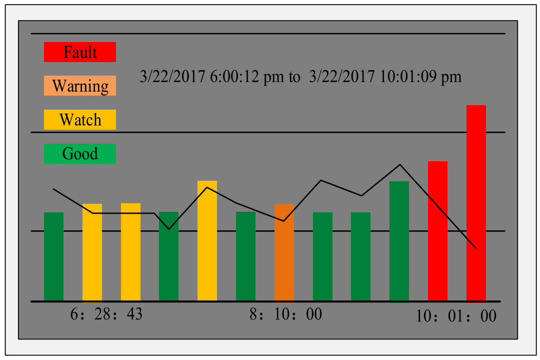

FIGURE 5. The diagnosis of the DEKG.

Figure 5 is the diagnosis of the DEKG, it warns nearly two hours and alerts three and a half hours in advance.

C. DISCUSSIONS

The traditional diagnosis is a process of the FAF. It outputs the superficial understanding of a problem at the time it occurs. The DEKG method can grasp the depths of the problem using diagnostic methods based on big data. It greatly reduces the dependency on experts' experience and raises the degree of diagnostic automation. Supplemented by visualization representation, it provides a successful transition to the PAP, and helps to optimize the manufacturing process.

The DEKG system also has some limitations. First, it collects and analyzes the duration of actions, which, probably, makes diagnosis too partial. Second, it only predicts the type of failure, but ignores its severity. This leads to priority issues when making the decision on fault recovery. In the future, the DEKG can be illustrated in multiple dimensions and multiple levels, and combine more advanced big data analytical methods to overall improve the performance of diagnosis.

VI. ISSUE AND CHALLENGES

Starting from data acquisition, there are huge differences between the previous diagnosis methods and big data methods. The approaches of the former methods process small amounts of monitoring data sequentially with human interventions. In the context of big data, it is not realistic to rely on traditional diagnosis to monitor and predict devices' health conditions. Fault diagnosis based on big data is able to satisfy diagnostic demands, including processing and responding to massive data in real time, extracting fault features, and predicting failures automatically and effectively. However, it is still in its initial stage of development, having many issues and challenges.

A. DATA QUALITY AND COST BALANCE

With respect to data acquisition and data processing, data is the cornerstone of fault diagnosis. The quality of data directly affects the process of extracting fault features, and influences the validity and authenticity of results. There are a lot of mature technologies to gather massive distributed operating data [86]. But, first, data formats are different when using various standards when collecting it. Moreover, quality problems, such as data loss, noise, inconsistency [87], and imbalance [88], should not be ignored. Also, certain costs are required to layout some diagnostic facilities and implement the process of data collection, storage and uploading [22], [89]. The purpose of fault diagnosis is to reduce losses caused by downtimes to ensure vested interests or increase revenues. If diagnostic costs are higher than losses due to failures, fault diagnosis loses its meaning to some extent. Therefore, the question of how to obtain raw data of high quality, and at the same time ensure that the collection costs are acceptable, must be considered.

B. METHOD SELECTION AND APPLICATION PROBLEMS

With respect to diagnostic methods, self-diagnosis without human interventions also has some problems, such as the time costs of the training model, and the reliability of the trained model. Some common questions can be easily solved using expert experience, when it is a waste to train a new model. Accordingly, interactions with humans constitute another issue to be considered [90]. In order to extract valuable fault information more efficiently and reduce the cost, it is necessary to consider the optimum combination of diagnostic methods introduced above in the stage of design. Excellent diagnosis can optimize the production process and guarantee the optimum operating condition of equipment, to realize safe production and the promotion of benefits. Furthermore, the security of big data methods has always been the focus of attention [91], [92]. It is an obstacle to extend and implement fault diagnosis systems.

C. DEEP UTILIZATION

With respect to diagnostic results, for one thing, most of fault diagnosis methods remain at the theoretical and experimental stages. The effects of diagnosis and feasibility remain to be examined. One needs to seek opportunities to combine theory and practice, as the value of results can be more practical this way. For another thing, almost all diagnosis methods stop as results are obtained without considering further value mining. Feeding diagnosis back to the device producer, forming a closed-loop control to enhance production efficiency from the source and obtain more insight, is worth deep reflection. Moreover, most of diagnosis methods resolve superficial problems lacking the deep exploration of the essence of faults. Future research should focus more on dynamic characteristics and correlations within large complex equipment systems, and explain the causality of failures from the surface to the center.

VII. CONCLUSIONS

In this paper, we first comprehensively review and study the development of fault diagnosis. It can be divided into knowledge-driven, data-driven and value-driven diagnosis based on different driving factors. From the perspective of applicability in the big data era, we point out the limitations of previous diagnostic methods, including the first two approaches. As the data volume increases exponentially, hidden important values lead researchers to explore more advanced and effective approaches to predict and provide diagnosis of device failures. We classify fault diagnosis based on industrial big data as value-driven methods. They focus on discovering and deep understanding of diagnostic values hidden in massive multi-sourced heterogeneous data. Further, we discuss an application of big data for fault diagnosis in manufacturing. A novel diagnostic idea called the DEKG is introduced, which transforms a diagnostic strategy form the FAF into the PAP by monitoring the devices' heartbeats. In addition, we analyze and summarize existing issues and challenges in the age of big data from the aspects of data quality and cost balance, method selection, application problems, and deep utilization. It would be valuable to fuse the advantages of different diagnostic methods and take use of the extracted fault futures further, for example, forming a closed loop diagnostic control system and making maintenance decisions automatically. This provides thoughts for future studies.

REFERENCES

[1] J. Wan, D. Zhang, S. Zhao, L. T. Yang, and J. Lloret, "Context-aware vehicular cyber-physical systems with cloud support: Architecture, challenges, and solutions," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 106–113, Aug. 2014.
[2] X. Li, D. Li, J. Wan, A. V. Vasilakos, C. F. Lai, and S. Wang, "A review of industrial wireless networks in the context of industry 4.0," *Wireless Netw.*, vol. 23, no. 1, pp. 23–41, 2017.
[3] Z. Shu, J. Wan, D. Zhang, and D. Li, "Cloud-integrated cyber-physical systems for complex industrial applications," *Mobile Netw. Appl.*, vol. 21, no. 5, pp. 865–878, 2016.
[4] J. Wan, M. K. Khan, M. Qiu, and D. Zhang, "Cloud-assisted industrial systems and applications," *Mobile Netw. Appl.*, vol. 21, no. 5, pp. 822–824, 2016.
[5] J. Wan, M. Yi, D. Li, C. Zhang, S. Wang, and K. Zhou, "Mobile services for customization manufacturing systems: An example of industry 4.0," *IEEE Access*, vol. 4, pp. 8977–8986, 2016.
[6] Gartner. (2001). *Big Data*. [Online]. Available: http://www.gartner.com/it-glossary/big-data
[7] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the Internet of Things: Literature review and challenges," *Int. J. Distrib. Sensor Netw.*, vol. 2015, Apr. 2015, Art. no. 431047, doi: 10.1155/2015/431047.
[8] S. Din, A. Ahmad, A. Paul, M. M. U. Rathore, and J. Gwanggil, "A cluster-based data fusion technique to analyze big data in wireless multi-sensor system," *IEEE Access*, vol. 5, pp. 5069–5083, 2017.
[9] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017, doi: 10.1109/TII.2017.2650204.
[10] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, Sep. 2015.
[11] X. Dai and Z. Gao, "From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2226–2238, Nov. 2013.
[12] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, Jul. 2014.
[13] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Apr. 2017.
[14] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 293–311, 2003.

[15] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part III: Process history based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 313–326, 2003.
[16] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 327–346, 2003.
[17] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.
[18] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part II: Fault diagnosis with knowledge-based and hybrid/active approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3768–3774, Jun. 2015.
[19] P. Henriquez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Review of automatic fault diagnosis systems using audio and vibration signals," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 642–652, May 2014.
[20] S. Wang, J. Wan, D. Zhang, D. Li, and C. Zhang, "Towards smart factory for industry 4.0: A self-organized multi-agent system with big data based feedback and coordination," *Comput. Netw.*, vol. 101, pp. 158–168, Jun. 2016.
[21] J. Wan, S. Z. Tang Shu, D. Li, and S. Wang, "Software-defined industrial Internet of Things in the context of industry 4.0," *IEEE Sensors J.*, vol. 16, no. 20, pp. 7373–7380, Sep. 2016.
[22] L. Kong, D. Zhang, Z. He, Q. Xiang, J. Wan, and M. Tao, "Embracing big data with compressive sensing: A green approach in industrial wireless networks," *IEEE Commun. Mag.*, vol. 54, no. 10, pp. 53–59, Oct. 2016.
[23] D. Zhang, J. Wan, C. H. Hsu, and A. Rayes, "Industrial technologies and applications for the Internet of Things," *Comput. Netw.*, vol. 101, pp. 1–4, Sep. 2016.
[24] D. Zhang, Z. He, Y. Qian, J. Wan, D. Li, and S. Zhao, "Revisiting unknown RFID tag identification in large-scale Internet of Things," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 24–29, May 2016.
[25] V. Mayer-Schonberger and K. Cukier, *BIG Data: A Revolution That Will Transform how we Live, Work, and Think*. London, U.K.: John Murray, 2013, pp. 1–24.
[26] K. Tidriri, N. Chatti, S. Verron, and T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges," *Annu. Rev. Control*, vol. 42, pp. 63–81, Sep. 2016.
[27] A. Alzghoul, B. Backe, M. Löfstrand, A. Byström, and B. Liljedahl, "Comparing a knowledge-based and a data-driven method in querying data streams for system fault detection: A hydraulic drive system application," *Comput. Ind.*, vol. 65, no. 8, pp. 1126–1135, 2014.
[28] H. Ma and G. Yang, "Observer-based fault diagnosis for a class of nonlinear multiple input multiple output uncertain stochastic systems using B-spline expansions," *IET Control Theory Appl.*, vol. 5, no. 1, pp. 173–187, 2011.
[29] W. Chen, W.-T. Chen, M. Saif, M.-F. Li, and H. Wu, "Simultaneous fault isolation and estimation of lithium-ion batteries via synthesized design of Luenberger and learning observers," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 1, pp. 290–298, Jan. 2014.
[30] G. H. B. Foo, X. Zhang, and D. M. Vilathgamuwa, "A sensor fault detection and isolation method in interior permanent-magnet synchronous motor drives based on an extended Kalman filter," *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3485–3495, Aug. 2013.
[31] M. Sepasi and F. Sassani, "On-line fault diagnosis of hydraulic systems using unscented Kalman filter," *Int. J. Control, Autom. Syst.*, vol. 8, no. 1, pp. 149–156, 2010.
[32] T. Jiang, K. Khorasani, and S. Tafazoli, "Parameter estimation-based fault detection, isolation and recovery for nonlinear satellite models," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 4, pp. 799–808, Jul. 2008.
[33] S. Zhai, W. Wang, and H. Ye, "Fault diagnosis based on parameter estimation in closed-loop systems," *IET Control Theory Appl.*, vol. 9, no. 7, pp. 1146–1153, 2015.
[34] M. Zhong, Y. Song, and S. X. Ding, "Parity space-based fault detection for linear discrete time-varying systems with unknown input," *Automatica*, vol. 59, no. 1, pp. 120–126, 2015.
[35] H. M. Odendaal and T. Jones, "Actuator fault detection and isolation: An optimised parity space approach," *Control Eng. Pract.*, vol. 26, no. 1, pp. 222–232, 2014.
[36] M. Ling, H. Yuan, Q. Wu, B. Tian, H. Yuan, and C. Huang, "Research of A fault tree-based real-time diagnosis and monitoring platform," in *Proc. IEEE 9th Conf. Ind. Electron. Appl. (ICIEA)*, Hangzhou, China, Jun. 2014, pp. 1893–1897.
[37] Y.-K. Liu, G.-H. Wu, C.-L. Xie, Z.-Y. Duan, M.-J. Peng, and M.-K. Li, "A fault diagnosis method based on signed directed graph and matrix for nuclear power plants," *Nucl. Eng. Design*, vol. 297, pp. 166–174, Sep. 2016.
[38] L. Wang *et al.*, "Knowledge representation and general Petri net models for power grid fault diagnosis," *IET Generat. Transmiss. Distrib.*, vol. 9, no. 9, pp. 866–873, 2015.
[39] N. Zhao, S. Li, Y. Cao, and H. Meng, "Remote intelligent expert system for operation state of marine gas turbine engine," in *Proc. 11th World Congr. Intell. Control Autom. (WCICA)*, vol. 4. Shenyang, China, Jun./Jul. 2015, pp. 3210–3215.
[40] P. Huo, S. Bai, Y. Ge, Y. Zhang, and H. Zhang, "Research on expert diagnosis system for mechanical fault of high voltage circuit breaker based on fuzzy matrix and neural network technology," in *Proc. Int. Conf. Condition Monitor. Diagnosis (CMD)*, Xi'an, China, Sep. 2016, pp. 134–138.
[41] Y. Zhi-Ling, W. Bin, D. Xing-Hui, and L. Hao, "Expert system of fault diagnosis for gear box in wind turbine," *Syst. Eng. Procedia*, vol. 4, pp. 189–195, Sep. 2012.
[42] B. Cai, Y. Zhao, H. Liu, and M. Xie, "A data-driven fault diagnosis methodology in three-phase inverters for PMSM drive systems," *IEEE Trans. Power Electron.*, vol. 32, no. 7, pp. 5590–5600, Jul. 2017.
[43] J. Chen *et al.*, "Wavelet transform based on inner product in fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vols. 70–71, pp. 1–35, Sep. 2016.
[44] J. Yan and L. Lu, "Improved Hilbert–Huang transform based weak signal detection methodology and its application on incipient fault diagnosis and ECG signal analysis," *Signal Process.*, vol. 98, pp. 74–87, Apr. 2014.
[45] M. Grbovic, W. Li, P. Xu, A. K. Usadi, L. Song, and S. Vucetic, "Decentralized fault detection and diagnosis via sparse PCA based decomposition and maximum entropy decision fusion," *J. Process Control*, vol. 22, no. 4, pp. 738–750, 2012.
[46] S. Yin, X. Zhu, and O. Kaynak, "Improved PLS focused on key-performance-indicator-related fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1651–1658, Mar. 2015.
[47] H. A. Talebi and K. Khorasani, "A neural network-based multiplicative actuator fault detection and isolation of nonlinear systems," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 3, pp. 842–851, Apr. 2013.
[48] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, no. 5, pp. 1–15, 2014.
[49] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.
[50] Z. Yin and J. Hou, "Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes," *Neurocomputing*, vol. 174, pp. 643–650, Jan. 2016.
[51] R. Ma, J. Fu, and T. Chai, "Dwell-time-based observers design for unknown inputs switched linear systems without requiring strong detectability of subsystems," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 4215–4221, Aug. 2017, doi: 10.1109/TAC.2017.2683065.
[52] A. E. Rundell, A. Chakrabarty, M. Corless, G. T. Buzzard, and S. H. Zak, "State and unknown input observers for nonlinear systems with bounded exogenous inputs," *IEEE Trans. Autom. Control*, to be published, doi: 10.1109/TAC.2017.2681520.
[53] C.-L. Dong, Q. Zhang, and S.-C. Geng, "A modeling and probabilistic reasoning method of dynamic uncertain causality graph for industrial fault diagnosis," *Int. J. Autom. Comput.*, vol. 11, no. 3, pp. 288–298, 2014.
[54] X. Wang, C. Mahulea, and M. Silva, "Diagnosis of time Petri nets using fault diagnosis graph," *IEEE Trans. Autom. Control*, vol. 60, no. 9, pp. 2321–2335, Sep. 2015.
[55] Y. Wang, G. Xu, L. Liang, and K. Jiang, "Detection of weak transient signals based on wavelet packet transform and manifold learning for rolling element bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 54, pp. 259–276, Sep. 2015.
[56] Y. Shatnawi and M. Al-khassaweneh, "Fault diagnosis in internal combustion engines using extension neural network," *IEEE Trans. Ind. Electron.*, vol. 61, no. 3, pp. 1434–1443, Mar. 2014.

[57] X. Zhang, B. Wang, and X. Chen, "Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine," *Knowl.-Based Syst.*, vol. 89, pp. 56–85, Nov. 2015.
[58] L. Shu, Y. Chen, Z. Huo, N. Bergmann, and L. Wang, "When mobile crowd sensing meets traditional industry," *IEEE Access*, vol. 5, pp. 15300–15307, Jan. 2017, doi: 10.1109/ACCESS.2017.2657820.
[59] S. Wang, J. Wan, D. Li, and C. Zhang, "Implementing smart factory of industrie 4.0: An Outlook," *Int. J. Distrib. Sensor Netw.*, vol. 2016, Apr. 2016, Art. no. 3159805.
[60] A. Labrinidis, Y. Papakonstantinou, J. M. Patel, and R. Ramakrishnan, "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, 2014.
[61] S. Madden, "From databases to big data," *IEEE Internet Comput.*, vol. 16, no. 3, pp. 4–6, Apr. 2012.
[62] C. Yang, C. Liu, X. Zhang, S. Nepal, and J. Chen, "A time efficient approach for detecting errors in big sensor data on cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 2, pp. 329–339, Feb. 2015.
[63] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[64] A. Bahga and V. K. Madisetti, "Analyzing massive machine maintenance data in a computing cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 10, pp. 1831–1843, Oct. 2012.
[65] A. Kumar, R. Shankar, A. Choudhary, and L. S. Thakur, "A big data MapReduce framework for fault diagnosis in cloud-based manufacturing," *Int. J. Prod. Res.*, vol. 54, no. 23, pp. 7060–7073, 2016.
[66] L. Yang, S. Yao, Y. Teng, X. Yang, X. Li, and J. Ren, "Research on fault diagnosis of photovoltaic power generation system based on cloud computing data," in *Proc. Int. Conf. Smart City Syst. Eng. (ICSCSE)*, Hunan, China, Nov. 2016, pp. 594–597.
[67] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proc. Usenix Conf. Hot Topics Cloud Comput.*, Boston, MA, USA, Jun. 2010, pp. 1–10.
[68] W. Shi *et al.*, "An integrated data preprocessing framework based on apache spark for fault diagnosis of power grid equipment," *J. Signal Process. Syst.*, vol. 86, pp. 221–236, Mar. 2017.
[69] *Why Use Storm*. Accessed on Jul. 2017. [Online]. Available: http://storm.incubator.apache.org
[70] J. Wan *et al.*, "A manufacturing big data solution for active preventive maintenance," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2039–2047, Aug. 2017, doi: 10.1109/TII.2017.2670505.
[71] Z. Huo, M. Mukherjee, L. Shu, Y. Chen, and Z. Zhou, "Cloud-based data-intensive framework towards fault diagnosis in large-scale petrochemical plants," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Paphos, Cyprus, Sep. 2016, pp. 1080–1085.
[72] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[73] P. Tamilselvan and P. Wang, "Failure diagnosis using deep belief learning based health state classification," *Rel. Eng. Syst. Safety*, vol. 115, no. 7, pp. 124–135, 2013.
[74] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings," *Mech. Syst. Signal Process.*, vols. 72–73, no. 2, pp. 92–104, 2016.
[75] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mech. Syst. Signal Process.*, vols. 3–72, pp. 303–315, May 2016.
[76] T. P. Banerjee and S. Das, "Multi-sensor data fusion using support vector machine for motor fault detection," *Inf. Sci.*, vol. 217, no. 24, pp. 96–107, 2012.
[77] J. Xiong, Q. Zhang, G. Sun, X. Zhu, M. Liu, and Z. Li, "An information fusion fault diagnosis method based on dimensionless indicators with static discounting factor and KNN," *IEEE Sensor J.*, vol. 16, no. 7, pp. 2060–2069, Jul. 2016.
[78] W. Chen and A. M. Bazzi, "Logic-based methods for intelligent fault diagnosis and recovery in power electronics," *IEEE Trans. Power Electron.*, vol. 32, no. 7, pp. 5573–5589, Jul. 2017.
[79] H. Peng, J. Wang, M. J. Pérez-Jiménez, H. Wang, J. Shao, and T. Wang, "Fuzzy reasoning spiking neural P system for fault diagnosis," *Inf. Sci.*, vol. 235, no. 6, pp. 106–116, 2013.
[80] L. Bennacer, Y. Amirat, A. Chibani, A. Mellouk, and L. Ciavaglia, "Self-diagnosis technique for virtual private networks combining Bayesian networks and case-based reasoning," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 1, pp. 354–366, Jan. 2015.
[81] J. Xu, K. Sun, and L. Xu, "Data mining–based intelligent fault diagnostics for integrated system health management to avionics," *Proc. Inst. Mech. Eng. O, J. Risk Rel.*, vol. 229, no. 1, pp. 3–15, 2015.
[82] J. Huang, G. Chen, L. Shu, Q. Zhang, and X. Wu, "WSNs-based mechanical equipment state monitoring and fault diagnosis in China," *Int. J. Distrib. Sensor Netw.*, vol. 2015, Sep. 2015, Art. no. 528464, doi: 10.1155/2015/528464.
[83] D. Djurdjanovic, J. Lee, and J. Ni, "Watchdog agent—An infotronics-based prognostics approach for product performance degradation assessment and prediction," *Adv. Eng. Informat.*, vol. 17, nos. 3–4, pp. 109–125, 2003.
[84] *Machine Monitoring Applications in the Oil and Gas Industry*. Accessed on Mar. 2017. [Online]. Available: http://www.sensorcloud.com
[85] *ENVISION*. Accessed on Mar. 2017. [Online]. Available: http://www.beet.com
[86] L. Shu, M. Mukherjee, M. Pecht, N. Crespi, and S. N. Han, "Challenges and research issues of data management in IoT for large-scale petrochemical plants," *IEEE System J.*, to be published, doi: 10.1109/JSYST.2017.2700268.
[87] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Apr. 2017.
[88] M. Grbovic, W. Li, N. A. Subrahmanya, A. K. Usadi, and S. Vucetic, "Cold start approach for data-driven fault detection," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2264–2273, Nov. 2013.
[89] X. Ding, Y. Tian, and Y. Yu, "A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1232–1242, Apr. 2016.
[90] J. Yan, Q. Liu, W. Xu, D. T. Pham, and C. Ji, "Servitisation of fault diagnosis for mechanical equipment in cloud manufacturing," in *Proc. IEEE Int. Conf. Ind. Informat. (INDIN)*, Cambridge, UK, Jul. 2015, pp. 586–590.
[91] J. Li, F. Tao, Y. Cheng, and L. Zhao, "Big Data in product lifecycle management," *Internal J. Adv. Manuf. Technol.*, vol. 81, no. 1, pp. 667–684, 2015.
[92] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the Internet of Things: Perspectives and challenges," *Wireless Netw.*, vol. 20, no. 8, pp. 2481–2501, Nov. 2014.

YAN XU received the B.A. degree in information management and information system from Shaanxi Normal University, China, in 2015. She is currently pursuing the Ph.D. degree with the School of Business Administration, South China University of Technology, China. Her research interests include fault diagnosis, big data, and machine learning.

YANMING SUN is currently a Professor with the School of Business Administration, Guangzhou University, China. He has directed multiple research projects, including the National Natural Science Foundation of China. He has authored or co-authored many scientific papers. His research interests include information systems, synergetics, and service-oriented manufacturing.

JIAFU WAN is currently a Professor with the School of Mechanical and Automotive Engineering, South China University of Technology, China. He has directed over 14 research projects, including the National Natural Science Foundation of China, the High-Level Talent Project of Guangdong Province, and the Natural Science Foundation of Guangdong Province. Thus far, he has authored over 120 scientific papers, including 70+ SCI-indexed papers, 20+ IEEE Trans./Journal papers, nine ESI Highly Cited Papers, and three ESI Hot Papers. According to Google Scholar, his published work has been cited over 2900 times. His SCI other citations (sum of times cited without self-citations) reached 705 times according to Web of Science Core Collection. His research interests include industry 4.0, industrial wireless networks, cyber physical systems, Internet of Things, clouding computing, embedded systems, and industrial robotics.

XIAOLONG LIU received the Ph.D. degree in management science and engineering from the South China University of Technology, China, in 2011. Since 2003, he has been a Lecturer with the School of Business Administration, South China University of Technology. His research interests include big data analysis, project management, and industrial informatics.

ZHITING SONG received the B.A. degree in industrial engineering from East China Jiaotong University, China, in 2013. She is currently pursuing the Ph.D. degree with the School of Business Administration, South China University of Technology, China. Her research interests include service-oriented manufacturing information systems, complex systems, and cyber-physical systems.