IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 November 2022, accepted 16 December 2022, date of publication 26 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232526

TOPICAL REVIEW

A Comprehensive Review on Big Data for Industries: Challenges and Opportunities

SUPRIYA SARKER[1,2], MOHAMMAD SHAMSUL AREFIN[2,3], (Senior Member, IEEE), MD KOWSHER[4], TOUHID BHUIYAN[3], PRANAB KUMAR DHAR[2], AND OH-JIN KWON[5]

[1]Department of Computer Science, University of Memphis, Memphis, TN 38152, USA
[2]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh
[3]Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh
[4]Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[5]Department of Electrical Engineering, Sejong University, Gwangjin, Seoul 05006, South Korea

Corresponding author: Oh-Jin Kwon (ojkwon@sejong.ac.kr)

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) Grant by the Korea Government (MSIT) (Development of JPEG Systems Standard for Snack Culture Contents) under Grant 2020-0-00347.

ABSTRACT Technological advancements in large industries like power, minerals, and manufacturing are generating massive data every second. Big data techniques have opened up numerous opportunities to utilize massive datasets in several effective ways to improve the efficacy of related industries. This paper presents a review of big data technologies used in the power, mineral, and manufacturing industries for various purposes. We analyze the meta-data of the collected papers before reviewing and selecting papers by applying selection criteria and paper quality assessment strategy. Then we propose a taxonomy of big data application areas in the power, mineral, and manufacturing industries. We have studied current big data architectures and techniques implemented in industry sectors and have uncovered the big data research gaps in industry sectors. To address the gaps, we point out some relevant research questions and, to answer the questions, we make some future research recommendations that might explore interesting research ideas for building a big data-driven industry. As the careful use of big data benefits every other industry sector; hence, supportive big data frameworks need to be developed to speed up the big data analysis process. Proper multi-dimensional big data assessment is also needed to take into account for serving effective data analysis tasks. Industry automation is also heavily influenced by the proper utilization of big data. While an intelligent agent can make many processes and heavy production loads in the manufacturing industry, it can work in a risky environment such as mines efficiently. To train agents for working in a specific environment big data can be used.

INDEX TERMS Big data for industry, smart grid, power system, oil and gas industry, minerals, big data technology, manufacturing industry, big-data-driven industry.

I. INTRODUCTION

The headway of information communication technologies (ICTs), the internet of things (IoT), and the age of industry 4.0 have brought all industries very near to automation [1]. With these advancements, a massive amount of data is generated every second, and it becomes clear that data is the most significant factor in the age of big data [2]. Big data can be called intellectual petroleum for all socio-technical sectors [3]. Revolutionary changes have resulted over the last few years with the adoption of big data technologies in leading industries such as power, mineral, and manufacturing.

Through the utilization of the invaluable data, many industries are reframing their operations in processing and reshaping their business models [1]. According to a survey in 2012 by International Data Corporation Energy, 70% of US oil companies were unconcerned about the application of big data techniques in the oil and gas industries [4]. Another

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

recent survey of General Electric and Accenture found that 81% of oil and gas companies considered big data as the top priority [5]. Thus, big data applications are increasing in energy sectors creating significant opportunities in energy conservation, energy management, environmental protection, energy consumption, and generated production data [4].

Nowadays, the large industries are benefited from big data techniques integrating IoT applications, machine learning (ML), and data mining algorithms to learn about their consumers, markets, and business trends from their operational data (*e.g.,* transaction price, electricity sales, electricity consumption and customers data) [6]. Many smartphone applications [7] with the help of integrated sensors are involved in collecting household continuous power consumption data from customers. The customers' consumption patterns help the companies to make significant business decisions, offer beneficial consumption policies to customers, and load distributions of an area. The production data *e.g.,* power generation and voltage stability data in the power industry allows the continuous monitoring of the system and detection of faults, or anomalies. Production data in the manufacturing industry can identify defective products, machines, and tools as well.

However, as industries produce highly unstructured and large dimensional data from diverse sources, the accumulation of such massive data in a unified structure and utilization of it is very challenging. For instance, oil and gas companies collect 2D, 3D, and 4D geophysical seismic-generated unstructured, complex data with the help of data-gathering sensors in subterranean wells to monitor operational resources. By developing region-wise data sharing models, oil and gas industries accumulate environmental data such as geographical properties, and marine life details in the sea-beds. These data-centric models may help to develop precise environmental models using the combination of ML with predictive analytics, and then environmental data-sharing models further can be considered for drilling fluid selection [8]. Besides, environmental models can diminish the current risks of inefficient drilling fluids that increase drilling time and costs and harm the environment. Thus, oil and gas organizations can select the optimum drilling fluid for a region. Environmental data sharing models can also be incentivized for oil and gas operators. Due to the current drop in oil and gas prices, they may help oil and gas operators to reduce the cost of decommissioning wells using the insights obtained from these models [9].

Moreover, large industries must tremendously emphasize controlled conditions [10] for operation and production. Though big data techniques are assembling vigorously in most large industries by establishing successful projects, they demand improved operational efficiencies and optimization of processes. Big data techniques such as data mining, training of the system, interpretation through predictive analysis by applying neural networks, classification and clustering algorithms can enhance productivity and efficiency [11]. Moreover, the big companies must focus on improving costs with increased profit margins and reducing the natural dangers (associated with power generation and mineral extractions) by using different data management techniques [12]. In this review paper, we try to find answers to the following questions:

- Which research topics on big data technologies related to the industry have been addressed so far?
- What are the big data research gaps that exist in the industry sectors?
- What are the open research questions that come up with potential solutions to the gaps?
- What would be the future directions to bridge the gaps?

Therefore, the motivation of the research study is to discover the big data research gaps that exist in the industry sectors. The main contributions made in this research study can be summarized as follows:

- We study the state-of-the-art big data technologies that have been applied to the development of the power, mineral, and manufacturing industries.
- We investigate existing research gaps in the power, mineral, and manufacturing industries.
- We propose some open research questions to bridge existing big data research gaps in the power, mineral, and manufacturing industries.
- We recommend some potential future research directions to eliminate the gaps and promote a big-data-driven industry.

The remaining paper is organized as follows. In section III, a taxonomy of big data application areas in the power, mineral, and manufacturing industries is proposed. The state-of-the-art research in these industries is discussed in sections IV-A, IV-B, and IV-C, respectively. Open research challenges and future directions are discussed in section V. Finally, section VII concludes the paper.

II. REVIEW METHODOLOGY OF BIG DATA APPROACHES FOR INDUSTRY

In this section, we discuss the process of conducting a review of research papers on big data for the industry. For searching and collecting research papers, we analyze the meta-data of research papers. The process of research papers' meta-data analysis is illustrated in Fig. 1.

We divide the entire process into four steps – identification, screening, assessment, and inclusion. In the identification step, we search papers using keywords following the search strategy discussed in II-A. At first, we selected some good sources of high-quality papers such as IEEE digital library, SpringerLink, ACM digital library, Science Direct, etc. We have searched the selected sources separately and have found a total of 2033 papers from IEEE (843), SpringerLink (358), ACM (117), Science Direct (430), and others (285). We apply the paper selection criteria discussed in II-B and remove 17 repeated papers that were found in the multiple databases and then the remaining 2016 research papers go to the screening step. In this step, we go through the papers' titles and abstracts to examine if they are written

in English and related to big data for the industry. Thus, we remove 1708 papers and the rest of the 308 papers go to the research paper quality assessment step where they are filtered by the quality assessment process discussed in II-C and demonstrated in Fig. 3. Finally, we get 132 papers for review.

A. SEARCH STRATEGY

We search the research papers on big data for industry sectors published before 2022 in some good quality popular research paper databases such as IEEE digital library, ACM digital library, SpringerLink, Elsevier, Multidisciplinary Digital Publishing Institute (MDPI), Google Scholar, Willey, etc. We search in each database using specific keywords such as *((big data <AND> industry) <OR> (power big data) <OR> (minerals big data) <OR> (process industry data))*.

B. STUDY SELECTION AND INCLUSION CRITERIA

To filter good quality papers from the collected papers, we set some inclusion and exclusion criteria are shown in Fig. 2. The selection and inclusion criteria are as follows:

- The research paper published before 2022
- Research topic relevant to big data for industry
- The research paper is written in the English language
- The research paper contributed to addressing its formulated research questions
- The research paper has got a quality score greater than or equal to three in the paper quality assessment process demonstrated in Fig 3.

C. COLLECTED PAPER QUALITY ASSESSMENT

We set some quality assessment questions to evaluate the quality of a research paper. Each question is worth one point. If a paper can answer a quality assessment question completely, it gains one point. Partially answering, the paper gains 0.5 points; otherwise, it gains a point of zero. Finally, all the points achieved by a paper are summed up which is the quality score of the paper. In this process, the quality of each paper is evaluated. The quality assessment process is demonstrated in Fig 3. The paper quality assessment questions are as follows:

- Does the paper clearly state its aims?
- Does the paper make a substantial novel contribution(s) in the field of big data for the industry?
- Does the paper discuss big data challenge(s)?
- Does the paper able to answer the formulated research question(s)?

In Table. 1, a year-wise paper distribution of the finally selected papers is listed. The year-wise frequency of papers is demonstrated in Fig. 5. We distributed the papers into six groups based on the range of years. As big data is a current emerging topic, we found that there are less number of papers before 2011. So, we listed the older papers till 2014 in a group. After that every year, we found a large number of publications.

III. TAXONOMY OF BIG DATA APPROACHES FOR INDUSTRY

From the literature review, several research papers on big data technologies implemented in the industry sectors have been found. To review finally selected papers in an organized manner, we propose a taxonomy shown in Fig. 4. After collecting the industry-related papers, we realize that based on the industry domains, we can categorize the existing research papers on big data for the industry into three main categories – the power industry, the minerals industry, and the manufacturing industry. Research papers that discuss big data technologies to handle power data, various methods, algorithms to improve the performance of smart grid systems, etc. are included in the power industry category. The minerals industry category includes research papers that discuss oil and gas field data accumulation, seismic data digitization and visualization, geophysical pattern recognition, drilling field identification, and so on. The manufacturing industry category includes production big-data handling techniques, product and machine fault detection methods, product quality assurance methods, and so on. Each category is divided into multiple sub-categories depending on the application sub-areas in each industry domain.

The power industry is sub-categorized into (i) power data quality assessment; (ii) power data fusion and cleaning; (iii) distributed power data mining; (iv) power data communication, privacy, and security; (iv) power data analytics. Among all the sub-categories, power data analytics is the most diverse sub-category. Therefore, we divide power data analytics into renewable energy prediction, system monitoring, fault detection, and user and business analytics. The minerals industry is sub-categorized into (i) minerals data storage and resource management; (ii) minerals data processing; and (iii) minerals data analytics. Like power data analytics, the implementation of minerals data analytics is also diverse. So, minerals data analytics is divided into exploration, drilling and completion, reservoir management, production engineering, pipeline monitoring, and maintenance. The manufacturing industry is sub-categorized into (i) manufacturing data processing, security, and transmission; (ii) process state monitoring; (iii) product quality assessment. Based on the implementation, manufacturing data analytics is divided into production management, product anomaly detection, and supply chain management. Though we classify the applied methods into different categories to assist our study, the techniques applied in one industry domain can also be applied in another domain to an extent except for some special cases. So, at the end of the study, we emphasize industry domain-independent big data techniques and discuss the overall challenges and future research opportunities in the industry sectors.

In Table 2 the applied techniques and methods used in big data in the power, mineral, and manufacturing industry are listed. We have listed the market tools that are used in big data processing and visualization in Table 3. Fig. 6 and Fig. 7 depicted the percentage of big data techniques to acquire and

TABLE 1. Year-wise paper distribution.

Till 2014	2015	2016	2017	2018	2019-2021
Wang *et al.* [13]	Ak *et al.* [14]	Adrian *et al.* [15]	Alam *et al.* [16]	Cadei [17]	Desai *et al.* [8]
Wang *et al.* [18]	Alfaleh *et al.* [19]	Alguliyev *et al.* [20]	Rollins [21]	Chan *et al.* [22]	Tsanousa *et al.* [1]
Kourti *et al.* [23]	Allen *et al.* [24]	Crespino *et al.* [25]	Shah *et al.* [26]	Cheng *et al.* [27]	Patel *et al.* [4]
Yacoub *et al.* [28]	Baaziz *et al.* [29]	Roden *et al.* [30]	Balouji *et al.* [31]	Fan *et al.* [32]	Li *et al.* [33]
Djurdjanovic *et al.* [37]	Betz *et al.* [34]	Ren *et al.* [35]	Bauman *et al.* [36]	Govindan *et al.* [38]	Patel *et al.* [4]
García-Muñoz *et al.* [41]	Cai *et al.* [39]	Gupta *et al.* [40]	Sukapradja *et al.* [42]	Hutchinson *et al.* [43]	Islam *et al.* [7]
McDaniel *et al.* [44]	De Santis *et al.* [45]	Li *et al.* [46]	Bello *et al.* [47]	Joshi *et al.* [48]	Jaeckle *et al.* [49]
Flores-Cerrillo *et al.* [53]	Han *et al.* [50]	Li *et al.* [51]	Bin Mahfoodh *et al.* [54]	Khvostichenko *et al.* [55]	Baek *et al.* [52]
Baruah *et al.* [56]	Jena *et al.* [57]	Liu *et al.* [58]	De Santis *et al.* [59]	Zhang *et al.* [60]	Dev *et al.* [61]
Tanabe *et al.* [62]	Tang *et al.* [63]	Liu *et al.* [64]	Duffy *et al.* [65]	Wu *et al.* [6]	Dinis *et al.* [66]
Canetta *et al.* [67]	Johnston *et al.* [68]	Niño *et al.* [69]	Han *et al.* [70]	Wei *et al.* [71]	Kumar *et al.* [72]
Chelmis *et al.* [73]	Kar *et al.* [74]	Palmer *et al.* [75]	Hashemi *et al.* [76]	Rossi *et al.* [77]	Song *et al.* [78]
Fan *et al.* [79]	Li *et al.* [80]	Sheng *et al.* [81]	Layouni *et al.* [82]	Maidla *et al.* [83]	Lv *et al.* [84]
Akoum *et al.* [12]	Li *et al.* [85]	Sun *et al.* [86]	Liu *et al.* [2]	Malbasa *et al.* [87]	Shukla *et al.* [88]
De Francisci *et al.* [91]	Ma *et al.* [89]	Swetapadma *et al.* [92]	Papadopoulos *et al.* [93]	Murray *et al.* [9]	Yuanting *et al.* [90]
Huang *et al.* [94]	MacGregor *et al.* [95]	Wang *et al.* [96]	Sarapulov *et al.* [97]	Ockree *et al.* [98]	Zhang *et al.* [99]
Kezunovic *et al.* [100]	Mishra *et al.* [101]	Wang *et al.* [102]	Wang *et al.* [103]	Olneva *et al.* [11]	Stergiou *et al.* [104]
Seemann *et al.* [105]	Mohammadpoor *et al.* [5]	Yang *et al.* [106]	Zhu *et al.* [107]	Singh *et al.* [108]	
Huang *et al.* [109]	O'Donovan *et al.* [110]	Zhang *et al.* [111]	Zhang *et al.* [112]	Taleb *et al.* [113]	
Lin *et al.* [114]	Raphael *et al.* [115]		Siryani *et al.* [116]	Yuan *et al.* [117]	
Ren *et al.* [118]	Sousa *et al.* [119]		Tu *et al.* [3]	Zhang *et al.* [120]	
Tao *et al.* [121]	Zhao *et al.* [122]		Udegbe *et al.* [123]		
Wang *et al.* [124]			Vezzetti *et al.* [125]		
Wu *et al.* [10]			Wang *et al.* [126]		
Zhang *et al.* [127]			Wu *et al.* [128]		
Zhao *et al.* [129]			Xianglan *et al.* [130]		
			Xiao *et al.* [131]		
			Yao *et al.* [132]		

TABLE 2. Applied technologies for power, mineral, and manufacture big data.

Area	Applied Technologies and Methods
Big Data Assessment	HDFS, HBase, kafka, FTP server, relational database Oracle, NoSQL, outlier detection
Data fusion and cleaning	Fuzzy method, support vector machine (SVM), radial basis function (RBF), neural network (NN), random forests (RF), and multi-layer perceptron (MLP)
Distributed data mining	Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP), Hadoop and Spark, LeNet-5 and LSTM networks, Directed Acyclic Graph (DAG)
Renewable energy prediction	Multilayer perceptron neural networks, multi-objective genetic algorithm, extreme learning combined, kd-tree
Power System Monitoring and Fault Detection	Dissimilarity measures learning, clustering techniques, fuzzy set-based decision rule, decision tree algorithm, Kalman filtering, hierarchical clustering, extreme learning machine, SVM, AdaBoost, active learning, Apriori, AprioriTid, and AprioriHybrid
Load forecasting	Neural networks, autoregressive integrated moving averages, and autoregressive moving average models
Power User and Business Analytics	Apriori algorithm, data mining and ML algorithms
Minerals Data Processing	Hybrid CPU/GPU system, MPI and CUDA parallel technology
Exploration through Seismic Pattern	Hadoop, principal component analysis, self-organizing maps, k-means clustering, multivariant regression, topological data analysis (TDA)
Production Engineering	Three-node Spark big data platform, PCA for dimensionality reduction, gradient boosting decision tree, RF algorithm, bootstrapping algorithm, logistic regression, decision tree, and neural networks
Pipeline monitoring and maintenance	Pattern-adapted wavelets, artificial neural network and linear regression
Manufacture Data Processing, Security, and Transmission	Process scheduling techniques in cloud service
Manufacture State Monitoring	Polynomial regression model, sequential clustering, hidden markov models, regression model
Product Quality Assessment	PCA, Projection to Latent Structures/Partial Least Squares (PLS)
Manufacture Data Analytics	Scheduling algorithms, IoT, and cloud computing in cloud manufacturing service platform for management Fast-AnoGAN, YOLOv3 for anomaly detection from the product image, Gaussian Restricted Boltzmann Machine for product quality inspection, RFID-enabled sensors for the collection of data in supply chain management

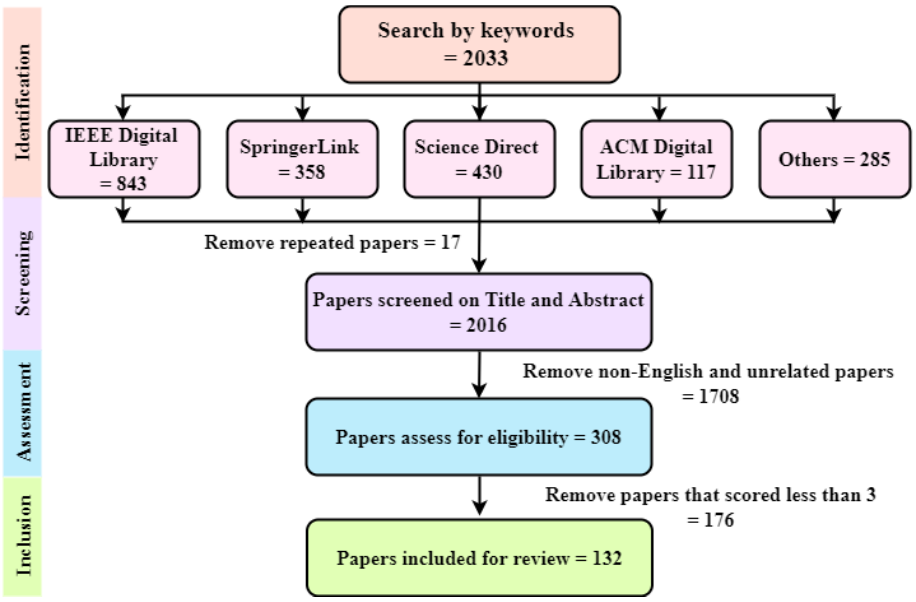

FIGURE 1. Meta-data analysis of research papers related to big data for industry.

TABLE 3. Big data market tools.

Tool	Description
Tableau	It is an excellent data visualization software that provides graphic representations by connecting big data from a wide range of sources, thus makes data analytics and business intelligence tasks easier. There is no need to code; however, because of advanced feature set, non-technical users sometimes find it difficult to use.
Apache Hadoop	It is an open-source data analytics software and its storage component (e.g HDFS) provides high-throughput access to application data and its processing component parallelly process large data sets but costs significant processing power. However, haddop does not provide real-time processing or in-memory calculations.
Apache Spark	Apache Spark is another open-source utility that enables real-time data processing using the underlying hardware's RAM and in-memory calculation. However, because it requires large amounts of RAM, it is less cost-effective compared to hadoop.
Apache Kafka	It is a distributed event store and stream processing tool. It provides a unified, high-throughput, low-latency for handling real-time data feeds.
Zoho Analytics	Zoho Analytics is an affordable big data analytics tool with an intuitive user interface. It can be directly integrated with Customer relationship management (CRM), Human Resources (HR), and marketing automation applications. But it lacks in advanced features, so inappropriate for large organizations.
MongoDB	It is a NoSQL database that does not use SQL-based rows and columns. It is good for load balancing and does not need any server. It is easy to use but processing is slow.
Xplenty	It is a cloud-based low-code ETL data integration tool that process streamline data from heterogeneous sources. Xplenty integrates with other applications like Zendesk, Oracle, and Salesforce and able to automatically integrate the data from all other tools.

process industry data and ML and data mining techniques to analyze and apply the data, respectively.

IV. LITERATURE REVIEW

A. STATE-OF-THE-ART BIG DATA TECHNIQUES IN THE POWER INDUSTRY

1) POWER DATA QUALITY ASSESSMENT

High-quality data is the most essential requirement for the adaptation of big data technologies and data analytics [39]. The performance and accuracy of any method heavily depend on the quality of big data and can be severely affected by poor-quality data.

A commonly adopted technique power data quality assessment is to frame quality assessment techniques to collect and clean the dataset using big data frameworks. A separate big data quality assessment framework for real-time and historical electric power data can be efficient in the power system

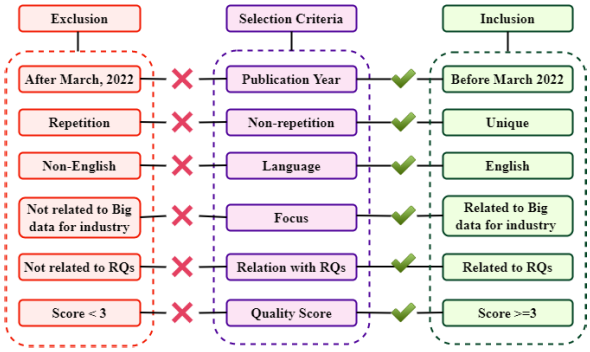

FIGURE 2. Research paper selection and inclusion criteria.

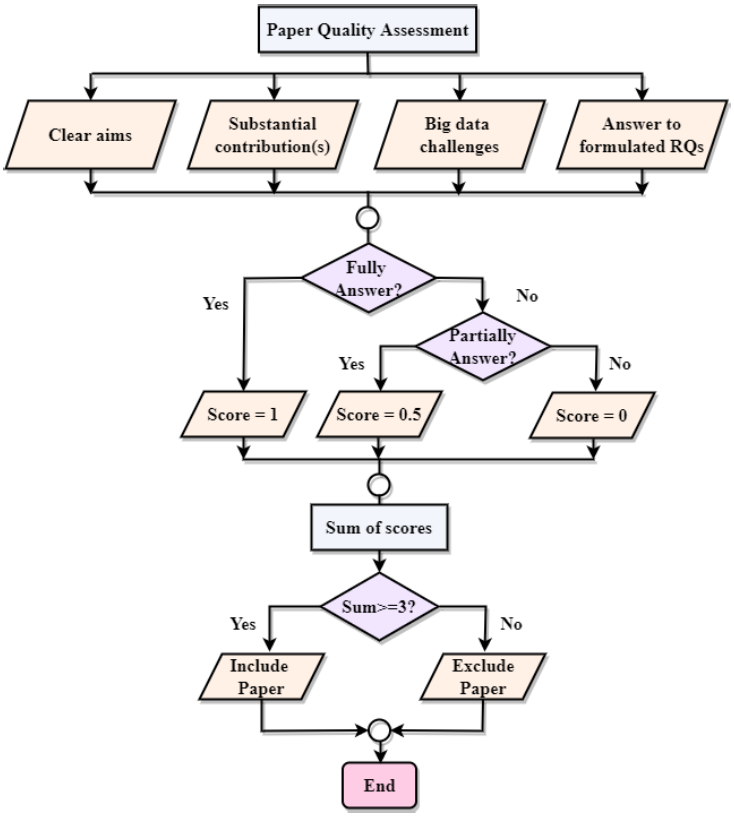

FIGURE 3. Illustration of paper quality assessment process.

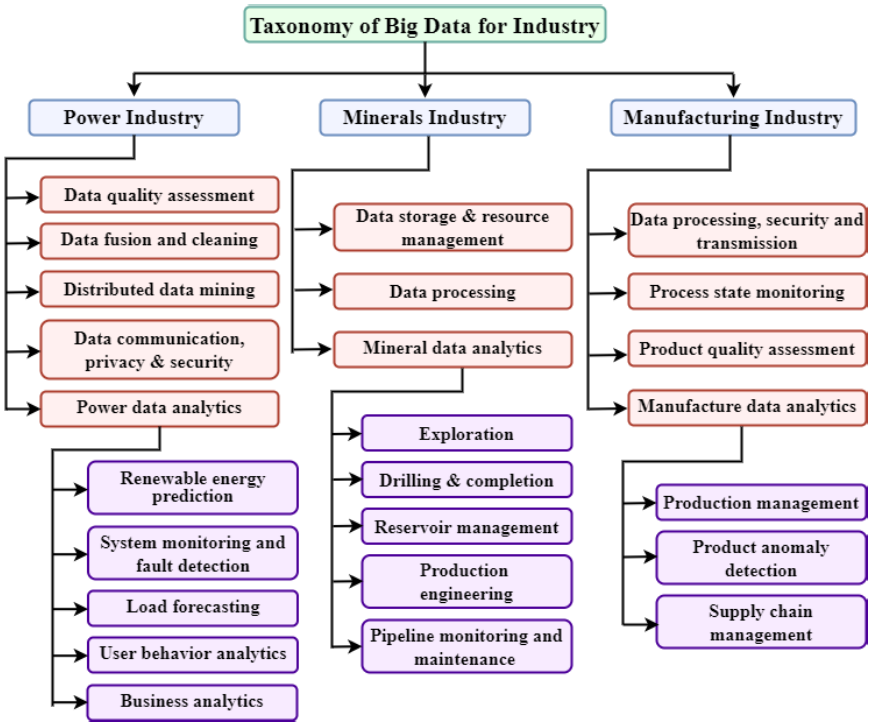

FIGURE 4. Taxonomy of big data for power, mineral, and manufacturing industry.

big data assessment system which is proposed in [2]. In this scenario, the power grid system can be separated into two separate sub-parts - headquarters and the provincial power grid. The power grid data can be acquired from provincial

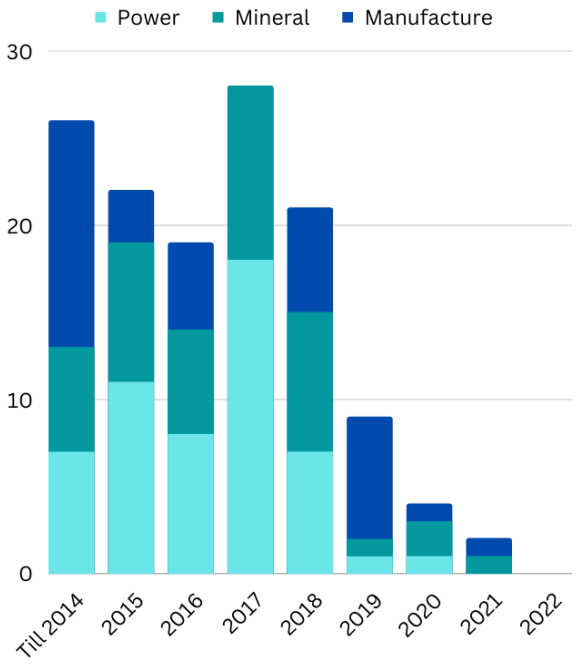

FIGURE 5. Year-wise distribution of reviewed papers.

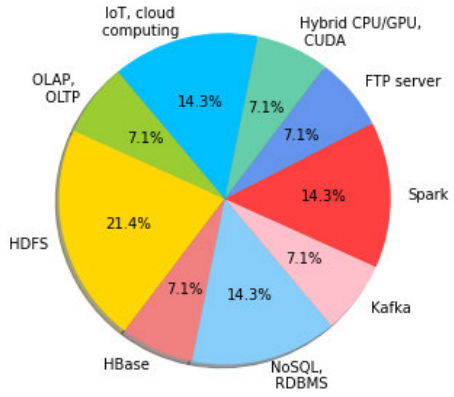

FIGURE 6. Big data technologies for data acquisition.

to headquarters using Kafka, and then real-time data can be stored using HBase. History data can be collected by a socket connection established between headquarters and the provincial grid; thus, headquarter can get access to the FTP server of the provincial grid, and historical data can be downloaded and stored in HBase and HDFS. Besides, an integrated database environment helps to store data that includes a relational database Oracle, a NoSQL database, HBase, along with a distributed file system HDFS. Before determining the type of assessment, redundant blank lines and blanks have been eliminated. The data quality assessment strategy determines if the data needs to be cleaned. For data cleaning, there are various data cleaning methods such as outlier detection. The data assessment strategy can be varied depending on data assessment types such as subjective and objective.

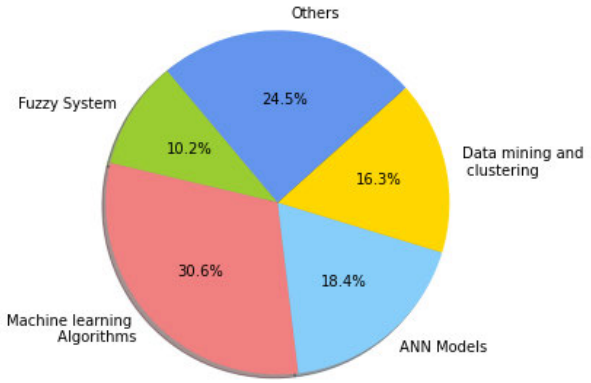

FIGURE 7. ML and data mining for big data processing and application.

However, an ideal assessment method must have the techniques to convert unstructured and semi-structured data to structured data and apply advanced techniques for extraction of meaningful features of data [113].

In addition to that, specific evaluation criteria for the power data quality assessment method must be defined. A clear process for selecting a quality assessment strategy for a certain domain was also ignored so far as well. The proficiency of such techniques must be proven by quality evaluation of power data through the practical experimental assessment of power grid systems.

2) DISTRIBUTED POWER DATA MINING

Efficient big data analysis systems must have powerful data processing and mining tools, algorithms, and platforms. Conventional data processing systems have been developed utilizing OLAP (Online Analytical Processing) system and OLTP (Online Transaction Processing) systems that are lacking standalone operational models. Hence, they proved insufficient for taking a long time in processing big data and unsatisfactory performance [126].

Enterprises are taking the help of data lakes to ingest data from on-premises, cloud, or edge-computing systems, to maintain full fidelity while storing and processing data, to analyze data using any language and applications. Distributed processing systems such as Hadoop and Spark, have brought a solution to the problem of traditional data pressing systems by incorporating deep learning frameworks. However, general-purpose mining algorithms are not serving well in discovering and utilization of specific purpose data such as industry data. So, building mining platforms *e.g.,* ([91], [132]) for domain-specific industry data has become a research focus of recent researchers.

Besides, for collecting and managing large power data virtually, IoT, cloud computing, and fog computing have proven highly efficient [104]. Stergiou and Psannis [104] proposed a framework for the energy consumption of industrial data centers across different heterogeneous machines. They embraced emerging reinforcement learning and federated

learning techniques. Recent research trends are using advanced distributed data processing frameworks such as Spark, Hadoop, YARN, and other frameworks with ML or deep learning models *e.g.,* such as [126] embraced deep learning distributed algorithms *e.g.,* LeNet-5 and LSTM networks, and implemented a Directed Acyclic Graph (DAG) to reduce the operational complexity and facilitate the reuse of components. However, how feature extraction, model learning, optimization, and insight have been utilized in the power industry is yet to explore.

3) POWER DATA FUSION AND CLEANING

Data fusion and cleaning is the foremost step in big data processing and analytics. Though data fusion and cleaning are very challenging because of the heterogeneity of big data. it is implemented by many big data platforms such as [122] where homogeneous data has been fused. Cluster and fuzzy methods are one of the efficient techniques for high-order heterogeneous data cleaning and fusion proposed in [109]. However, most of the methods were associated with a loss of data in the data cleaning process, which is not conducive to the mining of data in subsequent state assessments. To solve this problem, Lv et al. [84] proposed a data fusion method that can fuse multi-source heterogeneous grid data in text file format and store manual records of the grid into a unified format of data files. They proposed a data cleaning method based on ML using a support vector machine (SVM), radial basis function (RBF), neural network (NN), random forests (RF), and multi-layer perceptron (MLP).

4) POWER DATA COMMUNICATION, PRIVACY, AND SECURITY

The unrestricted nature of the smart grid enables the interconnection between the power grid and users [70], [80]. Access to huge numbers of users, extensive use of intelligent data acquisition, and wireless network transmission yield massive amounts of data. In power system transmission, each information level produces a huge amount of big datasets. With the surge of noisy data, conventional clustering algorithms are not efficient for widespread noisy data [80]. Moreover, the secure storage of data is one of the crucial challenges during the application of big data [80]. Including power system data, the privacy and security of user behavior data should be ensured. Besides, as the smart grid is an intelligent network that connects energy users' actions, reduces energy consumption and cost, and increases reliability by using advanced communications technology, suitable demand management is required to generate and transmit energy [79].

5) POWER DATA ANALYTICS

1) Renewable energy prediction
   Integration of smart grids with distributed renewable energy sources such as wind, solar, etc. causes increased challenges to the power industry. Though the operational capability of the power system to be fed into the energy grids has increased over time, the inconsistent nature of renewable energy may negatively impact the power supply [14]. So, the detection of possible renewable energy sources, and the prediction of the production of energy have vital importance. For prediction, artificial neural network outperforms most of the earlier methods. For instance, [14] predicted the prediction intervals using two separate methods to assess confidence in the prediction. In the first method, multilayer perceptron neural networks are trained with a multi-objective genetic algorithm, and in the second method, extreme learning is trained combined with kd-tree. The prediction of short-term wind on a real dataset of hourly wind speed measurements also provides us insight into the energy resources that will be required.
   However, though wind energy is one of the frontrunners in technological breakthroughs leading us towards more efficient power production systems, the drawbacks cannot be ignored. The establishment and maintenance of turbine and wind facilities are highly costly. This extreme expense is directly related to undeveloped technology. Through technological innovations, the reliability and energy output can be increased and system expenses can be reduced.
2) Power System Monitoring and Fault Detection
   Smart grid systems required continuous motoring of the power systems [87] and big data mining and ML algorithms facilitate the early predictions and forecasting regarding the states of power systems. Fault detection in power systems has great importance to avoid sudden power failure. The advancement of sensor technology allows collecting real-time information on power system health, operational status, and so on. In addition to the sensor-collected data, other data such as meteorological information proved useful to forecast power failure [58]. Therefore, fault detection in power systems is challenging as it involves multi-source heterogeneous data.
   Fault detection methods generally facilitate by clustering and classification techniques depending on the purposes. Smart grid fault detection can be considered a one-class classification problem if there is less variety in smart grid big data, otherwise, the classification could be multiclass classification. Classification can be done using various techniques while the fuzzy method, and decision tree algorithm are mostly used particularly for linear data and neural network mode can be used for non-leaner data. In the fuzzy method, a value is associated with every specific feature, and depending on the threshold the data can be classified. In the decision tree algorithm, based on the feature value decision is made on data instances, thus grid data can be classified as faulty or anomalous. Santisa et al. [45] and [59] considered it as a one-class classification problem and proposed a combined method of dissimilarity measures learning by evolutionary learning and clustering techniques. Then they analyzed the results

using a fuzzy set-based decision rule. They used power system operational data, Spatio-temporal data, physical components state data such as currents and voltages, weather conditions, etc. In [92] fault has been detected due to power swing within half cycle time using a decision tree. A decision tree algorithm was also applied for developing an intelligent relying on a transmission system by Jena and Samantaray [57]. They extracted 21 features from phasor measurement unit (PMU) data by applying Kalman filtering and feeding them into decision tree algorithms to provide the transmission line relaying decision. Papadopoulos et al. [93] applied a decision tree incorporating hierarchical clustering for detecting the dynamic behavior of power systems after an interruption and detected unstable non-synchronous groups.

Another purpose of fault detection is developing early warning systems. Reference [112] applied extreme learning machine (ELM) algorithm to develop an intelligent early-warning system for reliable online detection of risky events in the power system. Reference [96] analyzed the factors that affect transmission line galloping and proposed a bi-level classifier by applying SVM and AdaBoost.

In power systems transmission, a crucial characteristic of operation is maintaining voltage stability. Rapid decision-making has become challenging using massive amounts of data collected from geographically distributed locations applying traditional SCADA systems. A subset of ML *i.e.,* active learning has been proposed by Malbasa et al. [87] to predict the voltage stability of the power systems. The data-driven model supports online updates and offline training. Through active learning, unlabeled datasets are labeled automatically and data processing is done efficiently. By collecting the system operation data and meteorological conditions of the power system, Sheng et al. [81] adopted traditional association rule mining algorithms such as Apriori, AprioriTid, and AprioriHybrid incorporating a probabilistic graphical model to monitor the transformer state and predict failure.

3) Load forecasting

Load forecasting is vital for energy management, market demand analysis, and power system operation [60]. Load forecasting has been proposed by applying various techniques. However, centralized forecasting is challenging for a centralized power system due to weather diversity and load variation in different regions. Liu et al. [58] proposed distributed short-term load forecasting techniques from local weather data. The power system network is distributed in subnets based on optimal region partition and calculating similarity between vectors of influencing factors by cosine distance to select representative samples as training datasets. They used neural networks, autoregressive integrated moving averages, and autoregressive moving average models to separate subnets. Li et al. [85] proposed an extreme learning machine (ELM) by a wavelet-based ensemble scheme and to improve the performance of the model they integrated the Levenberg–Marquardt method. For feature selection and selecting input variables, they used conditional mutual information, and partial least squares regression has been implemented to combine all the particular forecasting results. However, feature selection or extraction is one of the most challenging tasks involved in ML models.

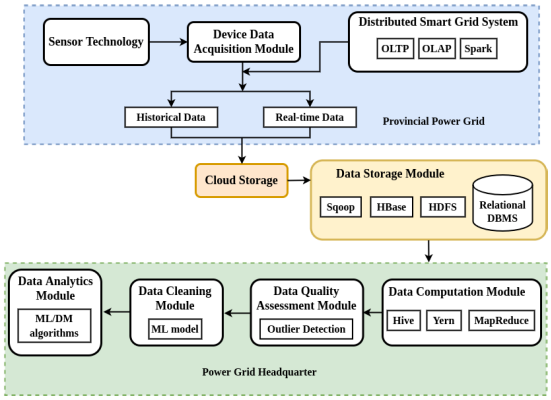

FIGURE 8. Big data techniques used for power grid data.

4) Power User and Business Analytics

Due to the improvement of living standards, the number of high-quality power appliances is increasing in houses. This creates an extra load demand during peak hours on the smart grid and is a challenge for safe operation [6]. The power load demand during different times of the day and locations widely varies. Therefore, to keep track of customer demand, power generation companies need to analyze load demand and ensure power supply without interruption. Big data technology can help to store residential power consumption data and discover user consumption behavior using data mining and ML algorithms. As a demonstration, Wu and Tan [6] proposed a big data storage framework for user power data and analyze the user behavior using the Apriori algorithm. Besides, Zhang et al. [111] emphasized on optimization of load demand response to manage home energy systems. This user data analytics assists power companies in making business decisions and updating policies for power users [26].

The complete process of power big data acquisition, processing, and application is depicted in Fig. 8 including the corresponding used technologies. A summary of some selected papers related to the power industry that applied big data technologies are listed in Tables 4 and 5. We have listed the major contributions, dataset, implementation and evaluation details, and limitations of selected papers. A category-wise list of the previous works in the power industry that embraced big data techniques is listed in Table. 6.

TABLE 4. Summary of some selected papers related to big data technology for the power industry.

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Taleb *et al.* [113]	An unstructured data quality assessment model	The authors used textual, media and web data or files, but did not mention a specific dataset name.	Data sampling has been done by BLB Bootstrap. Text mining methodology has used to extract significant information from the textual data and evaluate the data from the extracted features. Feature selection and mapping has been done using data quality dimension (DQD) metrics.	Performance of various data mining algorithms to assess data quality in industry domain have not been examined.
Liu *et al.* [2]	Proposed integrated storage using a relational database Oracle, a NoSQL database HBase and a distributed file system HDFS between headquarter and provision.	The authors proposed the framework for power data but did not mention specific dataset name.	Power grid data has been acquired from provincial to headquarters using Kafka and real-time data have been stored using HBase. Socket connection has been established between headquarter and provincial grid. For data assessment, the authors have considered subjective and objective data assessment type and applied various data cleaning methods.	The paper lacks the power grid data quality evaluation process of by the proposed a framework and a practical experimental assessment in the industry domain.
Wang *et al.* [126]	Proposed deep learning distributed algorithms	The authors proposed the framework for power data but did not mention a specific dataset name.	The authors extracted distributed unstructured text data using Spark; applied LeNet-5 and LSTM networks for data mining and to realize differential data analysis for power business data; implemented Directed Acyclic Graph (DAG) to reduce the operational complexity	The features extraction, neural model learning, optimization and insight that utilized in the power industry was not explained.
Huang *et al.* [109]	Proposed a fuzzy method to reduce data loss in power data cleaning process.	High-order heterogeneous data but did not mention a specific dataset name.	The proposed algorithm minimized distances between objects and centers of clusters in the feature spaces. The authors have derived fuzzy membership rules for the selection of objects and weights of features. The the number of the cluster has been estimated using the GXB validity index.	The authors have derived 9 fuzzy rules that may not present all the latent information in the heterogeneous data.
Lv *et al.* [84]	Proposed a data fusion method that can fuse multi-source heterogeneous grid data and store into a unified format of data files	They considered the stored files in the database as text files.	The authors proposed a data fusion method that can fuse multi-source heterogeneous power grid data from different systems and manual records of the grid into a unified format of data files. They proposed a data cleaning method based on machine learning using SVM, RBF, NN, RF, MLP.	They only used text files file as for fusion. The performance of fused data has not been evaluated in the application of the data.

B. STATE-OF-THE-ART BIG DATA TECHNOLOGIES IN THE MINERAL INDUSTRY

Big data has spread throughout diverse sub-areas of the minerals industry and has become one of the driving forces of the global economy and business. Minerals' big data may help to discover suitable oil and gas fields by nominal destruction of the environment; establish efficient mineral extraction systems; explore the relationship with customers, and benefit markets and opportunities. Therefore, research on the storage, analysis, and visualization of unstructured and semi-structure data has become the center of research interest and big data innovation in the mineral industry [130].

1) MINERALS DATA STORAGE AND RESOURCE MANAGEMENT

Typically, mineral systems involve an extensive amount of reports, geographical location maps, queries, official documents, and so on. Real-time monitoring and management of the exploration, mining, drilling, reservoir change, and utilization of associated management rights are impeded severely because of low efficient data collection, storage, and management systems [33]. The problem of inadequate natural energy resources compared to demand, inconsistent prices, environmental risks, and competition among other energy sources can be solved by proper utilization of data accumulated from diverse mineral sources [20].

Big data storage and management frameworks have demonstrated the ability to efficiently support large amounts of mineral data storage and management. Li et al. [33] investigated the architecture of a mineral resource management system and proposed a resource management system integrating big data and GIS technology. Besides, the proposed system also included various rights management such as mineral mining management, exploration management, GIS map management, geological exploration management, resource

TABLE 5. Summary of some selected papers related to big data technology for the power industry (cont).

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Ak *et al.* [14]	Predicted the Prediction intervals to assess confidence in the wind energy prediction	They used Canadian Climate Data-Environment Data which is a real dataset of hourly wind speed measurements.	The authors proposed multilayer perceptron neural networks with NSGA-II algorithm; and extreme learning combined with kd-tree.	As the authors predicted short-term wind, the neural network must hold the history of time-series data. So, they should consider a suitable NN model such as RNN, LSTM, and GRU.
Santisa *et al.* [45]	Proposed smart grid fault detection using One-Class classifier	They used ACEA historical heterogeneous data of period spans across 2009–2012 that included weather conditions, spatio-temporal data of currents and voltages.	The authors proposed a combined method of dissimilarity measures learning by evolutionary learning and clustering techniques; analyzed the results using a fuzzy set based decision rule.	The authors compared the proposed method with some UCI datasets that are not smart grid operational data. So, the comparison may not present the actual evaluation results.
Swetapadma *et al.* [92]	Proposed a decision tree-based framework for fault detection	The power system network has been simulated to generate fault in the power system by modifying its parameters. The fault breakers have been used to simulate various shunt faults.	To generate power fault, varying different fault parameters have been simulated using MATLAB software. Three-phase currents and voltages have been pre-processed using discrete Fourier transforms in the feature extraction step. The proposed model has been trained with a total of 20 no-fault instance and 240 power swings with fault samples.	Simulated data may not present the real power system fault data patterns. The authors have derived some rules to take the decision that is not sufficient to extract all the patterns in the power system.
Malbasa *et al.* [87]	Proposed active learning to predict voltage stability in the power system	The proposed method has been evaluated using synthesized data.	They proposed a data-driven ML models such as ANN, RF, SVM, Decision Trees that can be trained in both online and offline. They labeled the unlabeled voltage data using active learning automatically and efficiently. An experimental simulation has been conducted using MATLAB Neural Network Toolbox. The test network. The test network has consisted of 29 generators, 179 buses, 263 transmission lines, 42 shunts, and 104 loads.	Prediction time increases with increased training size in RF and SVM.

reserve management, resource database management, statistical analysis management of data, etc. [33]. In their proposed system, mineral assets data has been stored on the server-side of the responsible monitoring department of the government and can be accessed using an internal local area network connection. Mineral business enterprises can access the resources using the internet. However, the security of the system is a crucial issue that has been ignored. To protect invaluable mineral resource data from hacking, security systems need to be strengthened. Bello et al. [47] developed a big data architecture for automation of distributed production and downhole sensing data transmission, management, and visualization through the real-time reservoir and well monitoring applications. They further explored a web-based framework for data exchange between industries and business enterprises, monitoring, and interpretation [47]. Another integrated architectural model with big data business analytics and transaction data has been proposed by Alguliyev et al. [20].

2) MINERALS DATA PROCESSING

Due to progress in the petroleum and IoT sectors, extensive amounts of petroleum geoscience data have been growing in recent years and have spread to Peta Byte or Zeta Byte.

TABLE 6. Category-wise list of papers that implemented big data technology in the power industry.

Category	Related Paper
Power system operation	[100], [32]
Power data quality assessment	[2], [113], [39], [15]
Distributed mining	Mining with Spark [132] and combining Tensorflow [126], Social media big data streams mining [91]
Data fusion and cleaning	Fusion and storage of a homogeneous grid data [122], Multi-source heterogeneous data fusion with ML [84], Fuzzy joint clustering of heterogeneous data [109]
Data communication, privacy and security	Data processing, and secure big data technologies [80], [44], [129], Power network dispatching and planning [70], Smart grid communications [79]
Power data analytics	Renewable energy prediction [118], [14], [128], [106] power outage [86], [103], [36], voltage stability [87], [107], overcurrent [101] detection Anomaly detection [120], Load forecasting [58], Power fault [45], [59], power quality event [31] detection Transmission fault [92] detection Power swing fault, islanding [76], [16] detection [92], Transmission lines relaying decision making [57], [74] Unstable non-synchronous groups detection [93], Optimal risk analysis [112], Transmission line galloping [96], Transformer state parameter monitoring [81], HVAC system Monitoring with fuzzy logic and ANN [24], Smart meter operations with ML [116] Customers' energy consumption [6], demand response management [111] Analog meters data extraction [63] Power business analytics [26]

Petroleum data include seismic data (85%), well log data (6%), petroleum engineering data (5%), rock physical data (2%), and others (2%) [50]. So, it has become very challenging to handle a large amount of petroleum geoscience instantly using CPU only [89]. Han et al. applied a hybrid CPU/GPU system to process petroleum geoscience big data. MPI and CUDA parallel technology has been used to reduce computational costs. They used 8 GPU for the computation of actual seismic data and it took 48GB in 3.1 hours [50]. However, to run parallel computational tasks smoothly, the efficiency of computers is a crucial factor. For the computation of big data MPI, and CUDA require high-performance systems which are extremely costly. Therefore, alternative efficient computing technologies need to be proposed.

3) MINERALS DATA ANALYTICS

1) Exploration through Seismic Pattern
   Due to the advancement of geo-phones, a large amount of data is generated and captured at every moment of the fracture job [48]. Explanation of seismic data is essential for visual understanding. Many well-known oil and gas companies like ExxonMobil have utilized seismic visualization to detect or predict the distribution of fractures in tight reservoirs that improves streamline and well placement. Various sensors *e.g.,* hydrophones or geophones capture data from low-frequency waves caused by tectonic activities, and nowadays the well-developed oil and gas industry forms data-driven 3D visualization frameworks on topography, speed display, and depth imaging [8].
   Geographical and geophysical data aggregated from oil and gas-bearing basin and fields contributed a huge amount of data [11]. Olneva et al. [11] proposed a framework using Hadoop to analyze seismic datasets, extract significant geological features, characterize the reservoir and identify geological issues [48]. Multiple high-performance computers and advanced data analysis algorithms are involved to interpret seismic data. As an illustration, Roden proposed principal component analysis (PCA) and self-organizing maps (SOMs) to interpret shifts from a large amount of seismic data [30]. As PCA and SOMs are unsupervised ML algorithms, they can extract unknown hidden features from seismic data and help to understand the geology. By visualization, geological hidden features can be easily visible from a 2D color map.

As a practical realization, Olneva et al. [11] applied big data techniques to the database of the west Siberia petroleum basin, which is a great storage of diverse uncommon characteristics and hydrocarbon accumulation to discover new patterns in the distribution field and evolve novel exploration techniques. The authors (i) applied k-means clustering for 1D data, multivariant regression using drilling results from 5,000 wells data to demonstrate the regional structure of west Siberia through visualization of regional maps and charts, and (ii) build a training image sample of seismic events in the Achimov sequence from 3D seismic data and geological patterns from more than 40,000 sq. kilometer of 3D regional data.

Alfaleh et al. [19] applied topological data analysis (TDA) to analyze the shape of complex data, and discover clusters and their statistical importance for examining reservoir connectivity and compartmentalization. The application of TDA allows forecasting with high accuracy, new plans for growth, efficiency measurement, and optimization. Norne model [133] has been used to simulate the case study and inverted 4D time-lapse seismic data generated by reservoir simulations have been used.

2) Drilling and Completion

Analysis of a huge amount of data generated during the drilling process has a great impact on structuring the pipeline and further in detecting failures and ensuring safety during drilling operations. Manual processing of the real-time large amount of data prevents potential production. This process can be advanced by developing data-driven models that will intensify the execution of resources and increases production on wellheads [5]. Big data analytics and ML have become a current endeavor to improve the added value of drilling [40]. Through the utilization of wells' data collected by sensors, drilling models can be evaluated and geologic estimation of drilling procedures can be possible [65]. Besides, early detection of anomalies affects penetration and ensures avoidance of undesired events such as kicks, blowouts, etc. The quality of generated data is also necessary to evaluate to prevent the misuse of the drilling data and avoid future calamities resulting from wrong decision-making [83]. Through optimization, drilling operations can unite the logging [34].

Duffy et al. identified and standardized the best practices by incorporating an automated drilling state detection and monitoring service that speed up the production and boost the rig performance. The rigs' activity data can be transformed into meaningful performance patterns of crews after classification of data and integration of rigs' operational data. From identification of the most efficient crews, let them observe their working procedures and make other crews follow those procedures [65].

3) Reservoir Management

Big data provides an opportunity for reservoir engineers to monitor and sorting of reservoir simulation results. The explanation, system design, and prediction of the reservoir simulation parameters heavily rely on the analysis of stratigraphic rock [30], [115]. However, it is very difficult to digitalize 3D seismic data and estimate relative permeability parameters, and bottom hole pressure, which restricts the monitoring task of reservoir engineers [131]. Integration of cloud computing with big data provides optimizations of the reservoir parameters *e.g.,* gas lift, formation of water injection, water displacement spacing, and pattern in real-time [75]. Xiao and Sun applied a big data application model to predict the reservoir dynamics by assimilating all the data in the reservoir and production system and letting the reservoir engineers do continuous monitoring. The author later established relationships between different systems combining connected nodes and boundary conditions [131]. However, as there are so many reservoir parameters and varied patterns, a more efficient way of finding related features and boundary conditions should be investigated. To understand the induced and natural fracture features in the subsurface and their effect on the fluid flow and transport, Udegbe et al. [123] adopted a face detection approach using the cascade AdaBoost algorithm to discover patterns of fracture properties and gas shale production data classification using pattern recognition approaches from vectorized 1D image data. The performance of the proposed method was evaluated for hydraulically fractured wells. However, as ML models need to be trained with extracted features from production data, it is difficult to process and extract significant features to train the model in real-time.

4) Production Engineering

Production of oil and gas industries is heavily impacted by various damages *e.g.,* downhole casting, water injection, and life cycle of oil, gas, and water wells. Song and Zhou [78] adopted a method to predict casting damage by applying PCA for dimensionality reduction of data and then gradient boosting decision tree for supervised classification. They proposed a three-node Spark big data platform to collect test datasets from an oil field in mid-east China containing 446 wells data among which 352 are undamaged wells and 94 are casing damaged wells. In the data extraction process, they selected the 10 most significant parameters that are responsible for casting damage such as casting outer diameter, the thickness of the walls, the density of flow path between the near reservoir and the wellbore, sand layer bottom, sand layer top, etc. Then the dimensionality of parameters has been reduced by PCA to 5 dimensions and the risk of casting has been assessed using the Gradient Boosting Decision Tree algorithm. However, experimental data indicated that

some lacking in the collected dataset negatively impact the performance of the models, such as the absence of significant parameters and missing values, etc.

Ockree et al. [98] discussed many ML and data mining algorithms and their performance in classification and predicting well production. To demonstrate the working procedure of an ML model, they applied an RF algorithm to the wells' production dataset. For data sampling in the preprocessing step, they used a bootstrapping algorithm that randomly sampled data disregarding duplicate data. The bootstrapping algorithm has been used for replication of sample data and in this work the authors utilized approximately 1,000 replicate wells' data.

Cadei et al. developed a model that can forecast the H2S trespassing events and provide a rapid and broad solution by analyzing the main cause to early troubleshoot the fault [17]. They also discussed the effectiveness of various ML models as binary classifiers such as logistic regression, decision tree, and neural networks for forecasting the H2S trespassing events and they found that neural networks achieve high accuracy where logistic regression and decision tree improve the transparency of the forecasting. For feature extraction and training of the models, they used three types of abnormal events isolated from raw data, which were plant shutdown, gas sweetener shutdown, and PI recording failure.

5) Pipeline monitoring and maintenance

The pipeline is one of the important components in the oil and gas industries, while a huge percentage of petroleum is transported through the pipeline [82]. According to Canadian Energy Pipeline Association (CEPA) [134], the pipeline has been transporting 97% of natural gas and crude oil in Canada. Despite being considered one of the reliable and economical ways of transporting oil and gas, pipelines are frequently affected by various anomalies such as corrosion, cracks, and dents.

Various ML techniques have been used to detect anomalies in many research. Layouni et al. [82] discussed the numerical and non-numerical techniques to determine the defect lengths and depths and location of Metal-loss in the oil and gas pipeline and proposed a method to detect metal loss by applying pattern-adapted wavelets and two ML algorithms *i.e.,* artificial neural network and linear regression through investigating the magnetic flux leakage data collected from the scanning of oil and gas pipelines. They selected maximum magnitude, peak-to-peak distance, mean average, standard deviation, and integral of the normalized signal from 1,300 data items as defect depth heavily depends on these features.

A summary of some selected papers related to the minerals industry that applied big data technologies are listed in Tables 7 and 8. The complete process of mineral big data acquisition, processing, and application is depicted in Fig. 9 including the corresponding used technologies. A category-wise paper list of the previous works in the minerals industry that embraced big data techniques is listed in Table. 9.

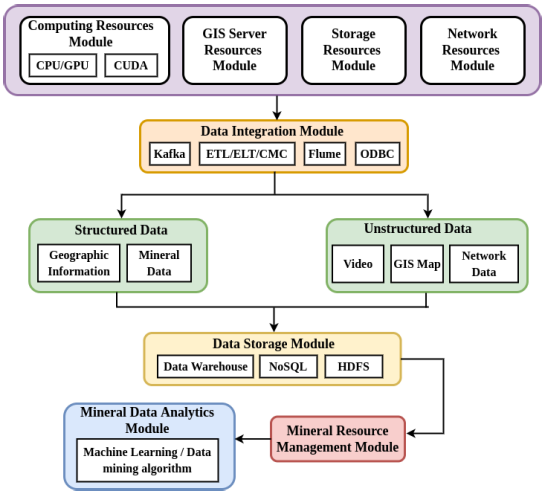

FIGURE 9. Big data techniques used for mineral data.

C. STATE-OF-THE-ART BIG DATA TECHNOLOGIES IN THE MANUFACTURING INDUSTRY

The highly available, efficient, and diverse sensors have aided the revolutionary transformation towards automation by supervising the functionality process, tools, machines, quality evaluation, fault prediction, and so on [1]. Through effective use of sensors and application of data fusion techniques, significant knowledge can be extracted and that can be further used for business growth.

1) MANUFACTURE DATA PROCESSING, SECURITY, AND TRANSMISSION

Manufacturing big data generally includes machine and tools-related data, operational data, business enterprise data, and external data. Wei et al. [71] developed the architecture of the service-oriented manufacturing data access model and discussed the various functionalities. The developed platform provides interfaces for various purposes such as web service, resetting the database, file upload, online filling of data, accessing the video, Web crawlers, etc. Responsible persons for data access configuration configured the access mode based on the requirements. Structured, unstructured, and semi-structured data can be transferred from the data center to the service center and vice versa with the help of a data service interface. In the service center, there were three modules *i.e.,* data collection units, data service units, and data distribution units. Finally, the service center is connected to a distributed big data storage system through a data bus channel. Liu et al. proposed a manufacturing model for the hydrostatic bearing system incorporating cloud with big data techniques [64]. The architecture and process scheduling techniques of the manufacturing industry in the cloud service

TABLE 7. Summary of selected papers related to big data technology for the minerals industry.

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Hum *et al.* [50]	Proposed a combined CPU/GPU system to process petroleum geoscience big data.	They used real seismic big data, but did not mention any dataset.	The authors embraced chaotic quantum particle swarm optimization and applied MRI and CUDA parallel processing to reduce the computational cost. They used 8 GPU for the computation of actual seismic data and it took 48GB in 3.1 hours. To test the algorithm, in 10 nodes 60 processes have been created.	For the computation of big data MPI and CUDA requires high-performance systems that are extremely costly.
Li *et al.* [33]	Proposed a resource management system integrating big data and GIS technology.	The architecture used geographic information, mineral data, video data, GIS map, and network data. The authors did not mention the dataset.	They proposed a system architecture to manage mineral resource data and introduced some right management through the proposed system. Mineral assets data has been stored on the server-side of the governmental monitoring department and can be accessed using an internal LAN.	To protect invaluable mineral resource data from hacking, security of the systems need to be strengthened.
Bello *et al.* [47]	Proposed an architecture to automate distributed production and downhole sensing data transmission, management, and visualization through the real-time reservoir and wells monitoring applications.	They used reservoir and wells production data but did not mention the dataset.	The proposed architecture has consisted of an input device, data collection, and aggregation; data storage, cleaning, and pre-processing; data retrieval, process, and analytical engine; data-driven decision support; intelligent control and information publish; graphical interface. Different wells data were converted to industry-standard data format and aggregated to a proxy server of the operational units. Later, data has been pushed to the system using IPSec tunnel. On the other hand, the system pulled data from wells through SCPI, and SFTP using Elastic Distributed Public and Private Cloud Data Center Landscape.	The security measures have not been explained.
Olneva *et al.* [11]	Applied big data techniques on the database to discover uncommon characteristics and hydrocarbon accumulation.	They used 40000 sq. kilometers of 3D regional data of west Siberia.	The authors (i) applied k-means clustering for 1D data, multivariant regression using drilling results from 5,000 wells data to demonstrate the regional structure of west Siberia through visualization of regional maps and charts, and (ii) build a training image sample of seismic events in the Achimov sequence from 3D seismic data and geological patterns from more than 40,000 sq kilometers of 3D regional data.	They applied k-means clustering only on 1D data.

platform have been investigated in the context of big data analytics in the paper [51].

It is crucial to ensure data security when capturing and transmitting data to and from the cloud computing environment. Data access should be controlled with the help of authenticated mechanisms so that data collected from one plant cannot be visible to people of other plants [69] and transmission should be secured by various powerful encryption algorithms.

2) MANUFACTURE STATE MONITORING

As the different machinery varies in their internal structure and working process, diverse techniques have been adopted to monitor their states, and performances and to detect faults [124]. Kumar et al. developed a health state of a cutting tool monitoring and estimation to aid automatic diagnostics, maintenance, and prediction using a polynomial regression model. The proposed model was built based on sequential clustering and applied to time-series sensor signals as unlabeled datasets [56], [72]. An inspection–replacement policy has been proposed by Zhang et al. [127] as a maintenance strategy for heterogeneous populations. In this policy, inspections are conducted at the early stage to find out and replace defective products, and at the later stage, a preventive replacement has been performed to avoid wear-out failures. Wang [18] applied machining processes using hidden Markov

TABLE 8. Summary of some selected papers related to big data technology for the minerals industry (cont).

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Xiao *et al.* [131]	Proposed a framework to predict the reservoir dynamics by assimilating all the data in the reservoir and production system	The authors used reservoir data, production data, and wellbore data.	The proposed application system has consisted of an ingestion system, geosystem, production data, wellbore data, surface gathering system, downhole pumps, and reservoir data. All the stored data has been transformed using hive, MAPR, and impala.	As there are so many reservoirs parameters and varied patterns, a more efficient way of finding related features and boundary conditions should be investigated.
Song *et al.* [78]	Predicted casting damage applying PCA for dimensionality reduction of data and then gradient boosting decision tree	The dataset contains 446 wells data among which 352 are undamaged wells and 94 are casing damaged wells.	They proposed a three-node Spark big data platform to collect test datasets from an oil field. They selected the 10 most significant parameters that are responsible for casting damage such as casting outer diameter, the thickness of the walls, the density of flow path between the near reservoir and the wellbore, sand layer bottom, sand layer top, etc.	Experimental data indicated that some lacking in the collected dataset negatively impact the performance of the models, such as the absence of significant parameters, and missing values. etc.
Ockree *et al.* [98]	They applied an RF algorithm to well production data for predicting well production.	They used well production datasets.	For data sampling in the preprocessing step, they used a bootstrapping algorithm that randomly sampled data disregarding duplicate data. The bootstrapping algorithm has been used for replication of sample data and in this work, the authors utilized approximately 1,000 replicate well data.	The negative effect of data duplication may have severe during training process of models.

TABLE 9. Category-wise list of papers that implemented big data technology in the mineral industry.

Category	Related Paper
Mineral Data Quality Maintenance	Heterogeneous wells data quality [54], oil field curation [73]
Petroleum Upstream Operation	Real-time ESP operation safety [40], Optimization [29] rod pump optimization [75], development planning [98]
Storage and resource management	The architecture of mineral resource management system with GIS [33], [20], MongoDB [117], Unconventional resource management [114], Well and reservoir management [47]
Mineral Data Processing	Drilling data automation [65], CPU and GPU parallel computing [50]
Mine Construction	Coal mine construction [130]
Mineral Data Analytics	Exploration [48], [119] Geographical feature extraction [11], Seismic interpretation [30], analysis [19], Drilling knowledge discovery [68], [83], optimization [43], Drilling cost optimization [77], [34] Production engineering [90], production allocation [21], well productivity [55], surveillance [115], Mineral extraction [131]
	Reservoirs modeling [42], management [105], evaluation [35], Upset and hazard prediction [17], [97], Casing damage prediction [78] Metal loss detection [82] Plants' health and safety [62] Business intelligence [12]

models (HMMs) for tool state detection. During tuning of the proposed model, feature vectors have extracted from variations of signals using a codebook that was built for vector quantization of the extracted features. Baek et al. developed a system to monitor the operational states of systems and detected faults through the identification of significant sensor signals using statistical variance and the Fisher criterion [52]. Wang et al. [13] emphasized a crucial issue of cost-effective interval time between critical level and the monitoring in condition-based maintenance. The authors investigated the relationship between the critical level and interval in condition monitoring to reduce cost and downtime and proposed a regression model based on the random coefficient growth model and assumed that coefficients followed the probability density distributions.

Besides, in the manufacturing industry manual, human efforts are reduced by intelligent agents. For example, Liu et al. [135] proposed a hierarchical structure model welding manufacturing system that introduced a leader following multi-agent robot to intelligentized welding manufacturing. The big data collected using IoT and sensor technologies helps to train the agent incorporating reinforcement learning.

3) PRODUCT QUALITY ASSESSMENT

Batch processing plays a vital role in various production processes and a massive amount of data is generated from this process. The data has a time dimension and a corresponding process variable and is collected during a batch, so the data has 3 dimensions. Product quality in certain batches is important to find faults in the product processing systems, assess product quality, and finally ensure sound production growth.

The assessment of the quality of the product by multivariate classification of collected data has been investigated by Garcia-Munoz [41]. To reduce the dimension of data, they used PCA, and after removing 12 outliers, a final PCA model of two components had been found that showed a separate spectrum for good quality and bad quality products. MacGregor et al. [95] also used latent variables regression models such as PCA and Projection to Latent Structures/Partial Least Squares (PLS) for analysis, monitoring, optimization, and control in the batch process of process industry. Through the projection of process data into low dimensional latent variable space, these dimensionality reduction methods can deal with highly correlated multivariate process data. A latent variable was also used to control the batch products in [53], product analysis and design in [49], for monitoring, testing, and performance measurement of products in [23], and optimization in [28].

4) MANUFACTURE DATA ANALYTICS

1) Production Management
   In the manufacturing industry, enhancement of efficacy of the system increases production growth and optimization of the process are the most crucial challenges in the age of globalization. Li et al. [51] followed a data analytics approach to discover an optimal strategy for workload management and proposed a novel scheduling algorithm based on a cloud manufacturing service platform [94]. Tao et al. [121] also proposed another similar cloud manufacturing service system based on IoT and cloud computing.
2) Product Anomaly Detection
   To increase the successful production in the manufacturing industry, anomaly or fault detection in the products has no alternatives. ML and DL techniques are currently widely used to detect faults in products. Jiang et al. [136] proposed a supervised anomaly detection model using YOLOv3, where anomalous products have been identified from balanced images. They also proposed a semi-supervised anomaly detection model using Fast-AnoGAN. The semi-supervised method generates new images using a trained WGAN-GP model. Zhang et al. [99] detected the anomalies in product quality inspection using Gaussian Restricted Boltzmann Machine. This model can handle high-dimensional and highly imbalanced distributions on product data.
3) Supply Chain Management
   Performance measurement plays a significant role in supply chain management of the manufacturing industry to measure the efficiency of a system. With the rapidly changing goals and limited personnel, performance measurement of individuals in a short time is very challenging for the organizations. High-level networking sensors such as RFID-enabled sensors offer the acquisition of real-time data for production logistic control in supply chain management. Supply chain visibility can enhance the predictive quality, monitor inventory, and boost customer service by tracking lot size, and distribution of production [67]. Large companies like Amazon and Walmart mine their clients' data for product promotion [61]. They visualize the supply and demand signal between retail stores and suppliers; optimize supply chain decisions.

A summary of some selected papers related to the manufacturing industry that applied big data technologies is listed in Table. 10 and 11. The complete process of manufacturing big data acquisition, processing, and application is depicted in Fig. 9 including the corresponding used technologies. A category-wise list of the previous works in the manufacturing industry that embraced big data techniques is listed in Table. 12.

V. BIG DATA OPEN RESEARCH CHALLENGES IN THE INDUSTRY

A. INDUSTRY DATA QUALITY ASSESSMENT

The quality of collected data heavily impacts the data analysis tasks. So, the assessment of the quality of the captured data is a crucial need that further calls for the necessity of data cleaning, and filtering for missing or invalid values. Currently, precise evaluation metrics for the assessment of big

TABLE 10. Summary of some selected papers related to big data technology for the manufacturing industry.

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Kumar *et al.* [72]	Proposed a system that can monitor the health state of cutting tools and estimate the remaining life.	They used unlabeled time-series datasets.	They proposed a hidden markov model (HMM) and polynomial regression model for monitoring the health state of cutting tools. The proposed model was developed based on sequential clustering of time-series sensor signals. The model performance was evaluated on the CNC machining testbed which was equipped with thrust-force and torque sensors to monitor drill bits.	The authors used hierarchical HMM to use an unlabeled dataset. unlabeled dataset. of the proposed model has not compared to other unsupervised techniques using unlabeled data or techniques that can label unlabeled data.
Wang *et al.* [18]	Applied HMMs for tool state detection	They collected the data using an accelerometer sensor mounted on the cutting tool holder attached to the turret.	The author has applied discrete wavelet transform to find wavelet coefficient and then extracted feature vectors from variations of signals using a codebook during the tuning of the model. The purpose of using the codebook was vector quantization of the extracted features. The testing of the model has been conducted on a 30hp CNC lathe and the signal was collected by an accelerometer.	The proposed method was not compared to other techniques.
Garcia-Munoz *et al.* [41]	Assess the quality of the product by multivariate classification of collected data	The data was considered as a three sets of variables (X, Y, Z) measured for each batch and prepared from chemical analysis of wet cake.	They used PCA and after removing 12 outliers, then a final PCA model of two components had been found that showed a separate spectrum for good quality and bad quality products.11 chemical variables and 1 weight has been collected from a chemical analysis of the wet cake drying process (initial condition matrix, Z). Then 10 process variables were found from the batch (process matrix, X) and 11 product quality variables were found at the end of the batch processing (quality matrix, Y).	PCA can only present the few most relevant components for analysis. So, linear dimensionality reduction methods have become obsolete for the extraction of rigorously emerging behavior from massive datasets.

data quality have not been defined and are lacking proper guidance. The current concept of big data has not defined the quality and criteria of big data. Big data quality depends on data type, format, features, domain, and so on [39], [113]. Moreover, big data frameworks that support rapid integration of big data from multiple industries are yet to be developed. Rapid quality assessment frameworks for massive amounts of industry data need to be developed.

Therefore, the research questions addressing the existing challenges for the quality assessment of industry data are as follows:

- How to define a big data quality assessment model including proper evaluation criteria?
- Can we develop a big data framework for the acquisition, storage, and processing of mixed structured data?
- How to examine the applicability of an assessment model for a particular industry?
- How can we develop appropriate big data infrastructure to accelerate the quality assessment process of the high-velocity industry data?
- How can we develop big data frameworks that would integrate and assess data quality across multiple industries when necessary?

B. INDUSTRY DATA FUSION AND CLEANING

Depending on the type of industry domain, the type and dimensions of the collected or stored dataset vary a lot. Therefore, data fusion frameworks need to be more flexible in type and dimension. Besides, many research works such as [84] focused on the frameworks to store the data, but were lacking clear discussion on how the ML techniques work on data cleaning and preparation for data mining techniques. The performance of the ML models has not been evaluated and optimized. The effective performance of heterogeneous data

TABLE 11. Summary of some selected papers related to big data technology for the manufacturing industry (cont).

Paper	Major Contributions	Used Dataset	Implementation & Evaluation	Limitations
Jiang [136]	Proposed a supervised anomaly detection model applying YOLOv3 on images and a semi-supervised anomaly detection model using Fast-AnoGAN.	They used image dataset of industrial production products.	The proposed model developed an ROI classifier to detect anomaly types. The semi-supervised model labeled the unbalanced image data set based on Fast-AnoGAN. To generate the image, they trained the WGAN-GP model and monitor the product score from the difference between generated image and the test image. The proposed models have been evaluated by a balanced and imbalanced dataset on the actual industrial production environment.	Performance with the generated images by the trained WGAN-GP model may mislead the proposed semi-supervised model.
Zhang [99]	Detected anomaly in product quality inspection using Gaussian Restricted Boltzmann Machine	High dimensional and highly imbalanced distributions on wine and cigarette product data	For pre-processing, they applied Zero-phase Component Analysis (ZCA) whitening techniques and integrated the free-energy function with the objective function and performed gradient compensation. As a training algorithm, Contrastive Divergence and Parallel Tempering was used. For performance metrics, the reconstruction error has been calculated.	They used ZCA which is very similar to PCA for feature extraction not reducing dimensionality. It may increase processing time and storage. More efficient feature extraction can be possible with an autoencoder.

TABLE 12. Category-wise list of papers that implemented big data technology in the manufacturing industry.

Sub-area	Related Paper
Data Access and Processing	Manufacture big data access platform [71] A hydrostatic bearing system with cloud and big data technology [64] Distributed manufacture data capture and integration [69] Latent variable modeling [95]
Process State Monitoring and Maintenance	Unified change detection [124] Health state estimation of cutting tools [72], [56], [18] Condition-based maintenance of heterogeneous products [127], estimate optimal critical level [13] Fault pattern extraction [52],
Product quality assessment	Macro-quality index based on customer satisfaction [46]
Big Data Analytics	Manufacture workflow management and scheduling on cloud [51], Capacity maintenance [66] Product manufacture anomaly detection [25] Data-driven cost estimation [22] Feedback and coordination system [102] Supply chain indicator [61], [38], [108], Smart manufacture [110], [27] and service [88] Product Lifecycle Management [125], [37]

fusion and storage solutions for practical industry systems has not been demonstrated.

Therefore, the research questions addressing the existing challenges in the fusion and cleaning of industry data are as follows:

- How can we develop big data fusion framework that is flexible to data type and storage format as necessary?
- How the model performance can be evaluated to facilitate further optimization?
- How to utilize the cleaned and fused heterogeneous multi-source industry data in solving current problems to prove the efficiency and accuracy of the data cleaning framework?

C. INDUSTRY DATA COMMUNICATION, PRIVACY, AND SECURITY

One of the great challenges of communication and transmission of industry data is data noise. Previous research on data

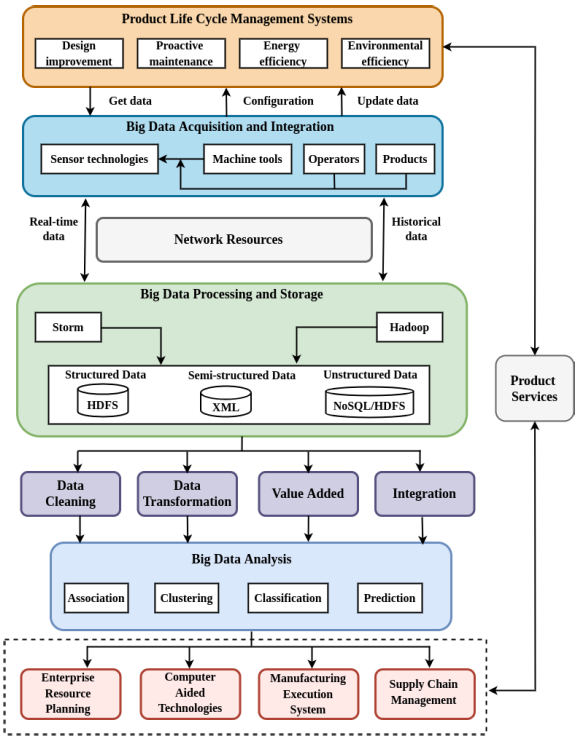

FIGURE 10. Big data techniques used for manufacturing data.

noise detection using traditional clustering algorithms is not much efficient for big data. Ensuring data security for the user and system is another big problem. When users' data has been utilized for various data analytics purposes, user privacy must be ensured.

Therefore, the research questions addressing the existing challenges of communication, privacy, and security of industry data are as follows:

- Can we develop a more efficient noise-protective big data architecture?
- How to detect noisy data using advanced data mining or ML algorithms?
- How advanced big data architecture can identify and manage the demand of users?
- How does big data infrastructure ensure the privacy of systems and users when using data for other purposes?

D. DISTRIBUTED INDUSTRY DATA MINING

Industry big data highly varies from one industry to another industry. For instance, the power and manufacturing industry produces lower-dimensional sequence data compared to high-dimensional geo-location maps used in the minerals industry. Therefore, generic distributed mining platforms will not serve industry-specific purposes. In addition to this, as ML and deep learning models are black-box type models, the decision-making process of those algorithms is very difficult to understand. As a result, optimization of the models becomes hard and is only limited to changes in neural models' hyperparameters.

Therefore, the research questions addressing the existing challenges in distributed mining of industry data are as follows:

- How to develop distributed domain-specific industry data mining platform along with big data technologies?
- How can we interpret neural models to ensure reusability and optimization?

E. PRODUCT QUALITY ASSESSMENT

Industry data is becoming highly non-linear and high-dimensional. Linear dimensionality reduction methods like PCA can only present the few most relevant components for analysis. So, linear dimensionality reduction methods have become obsolete for the extraction of rigorous emerging behavior from massive datasets. Besides, consumers' experiences regarding a product are very significant in measuring products' current quality and future improvements.

Therefore, the research questions addressing the existing challenges in product quality assessment from industry data are as follows:

- Which big data are significant for the understanding of the dynamic behavior of the manufacturing process and how can we capture the non-linear relationships of parameters?
- How can we develop the pipeline between product consumers' demand, experiences, and product assessment and quality enhancement?

F. INDUSTRY DATA ANALYTICS

1) SYSTEM MONITORING AND FAULT DETECTION

Industry systems need continuous monitoring to detect faults and it can be done through big data analysis. One of the most important and challenging tasks of data analysis is feature extraction. Almost in all of the previous research, feature extraction from industry data has been extensively done by applying ML or DL models. However, the causal correlation between the features remains hidden. The latent relationship may explore the reason for the fault or anomaly.

Therefore, the research question addressing the existing challenges in monitoring and identifying faults in the systems is as follows:

- How to discover the correlation/causation rules from the industry data?

2) MINERALS EXPLORATION AND DRILLING

Limited monitoring of fracture jobs has been done from seismic events at this time. The time lag between visualization and explanation of micro-seismic events restricts the capability to take real-time responses. Because of the narrow utilization of vast amounts of data, significant patterns have been remaining unexplored and hardly utilized for future task modeling and decision-making. Besides, practical implementation of most of these proposed ideas is yet to be applied in the mineral industry due to the involvement of natural risk and huge cost.

Therefore, the research questions addressing the existing challenges in the monitoring of seismic events for exploration and drilling are as follows:

- How to enhance technology for continuous monitoring of seismic events?
- How can we develop rapid visualization and interpretation of micro-seismic events and take real-time emergency measures?
- How to increase the use of seismic data to explore more latent features and thus offer an efficient data-driven solution to exploration?
- How can we implement the drilling ideas that avoid environmental harm and estimate cost?

3) RESERVOIR AND PRODUCTION MANAGEMENT

Datasets are a crucial component in the training of data mining and ML models. Improper or insufficient data hinders data-driven solutions. Therefore, more data should be provided or collected using big data technologies for the progress of mineral data analytics. Moreover, a proper framework should be developed for collecting good quality data and assessment of mineral industry data quality. Besides, data analytics requires ML models to be trained with production data and extract features. As system monitoring is a continuous real-time process, it is very challenging to collect data, pre-process, extract significant features, and train ML models in a real-time industry environment.

In addition, due to a lack of dataset, some authors [98] replicate the small dataset. But, replication or repetition and synthesis of datasets have severe drawbacks on the training of ML models. In this way, models learn the same features repetitively which demands unnecessary increased time and memory. We understand that models need a sufficient amount of data to be trained and for this reason, the necessity of duplication of an insufficient amount of data is justified; however, we cannot ignore the negative effect of duplication of data. Besides, more data can add to more varieties of patterns and can make trained models more efficient. Therefore, more operational and production original data from the oil and gas industries need to be collected for data analytics.

Forecasting or classification using neural network models is a black box type method that lacks enough transparency and interpretation. To serve the analysis, rapid and clear interpretation and visualization of results are crucial.

Therefore, the research questions addressing the existing challenges in monitoring mineral production are-

- How can we get enough good-quality datasets to use for various big data analytics purposes?
- How can we extract features and discover dynamically latent relationships between features, reservoir, and production system?

VI. DISCUSSION AND FUTURE RECOMMENDATIONS

The industry sectors have been revolutionized by big data technologies and have created unprecedented socio-economic development. In this review paper, we discussed the leading big data techniques involving industry and explored the underlying challenges to draw the attention of big data researchers and data analytics. In this paper, we recommend possible solutions to big data research challenges so that future researchers can work on those.

- Big data architecture has been proposed to gather industry data, and utilized it for several purposes. The missing value, noisy data, etc. in collected data impact negatively the later processing or use of this invaluable data. Therefore, data quality assessment is an urgent necessity. The currently proposed data quality assessment models are not precise and demand more guidance. A big data quality assessment model can be proposed and must define (i) data type, format, and domain; (ii) dimensions of data used for mapping the quality; (iii) quality metrics to consider; (iv) attribute or feature evaluation techniques; (v) data sampling strategies; (vi) assessment techniques. Non-relational databases e.g., NoSQL may be used to store heterogeneous data and an empirical big data quality assessment method can be investigated for individual industry domains. A weight coefficient must be defined for each assessment indicator. With big data frameworks like Spark, Storm, etc. to collect high volume data, high processing speed memory and algorithms have to be developed. State-of-the-art big data integration frameworks can be designed to collect multiple industry data and a single assessment model can be applied.
- Data cleaning and pre-processing are closely related to data quality assessment. Common data cleaning algorithms are used for typical industry data but do not consider high-dimensional data such as 3D or 4D seismic data, geographical location maps, and so on. So, an advanced big data framework for high dimensional mineral data storage and processing needs to be developed. The proposed data fusion and cleaning algorithms' performance needs to be measured in real-time industry environments. Advanced data cleaning algorithms may help to clean the noisy data during data transmission in the power industry as power data is susceptible to noise. Semi-supervised self-learning clustering algorithms can be developed to assist in the global data clustering of power system data. Besides, researchers should consider efficient noise-protective big data framework including big data sampling techniques to remove noise from smart grid-transmitted datasets.
- API and middleware can transfer energy data between smart grid and users [70]. By analyzing such data, users' demands can be identified and power systems can work to serve the increased demand. However, system and users' data privacy is a crucial challenge. Advanced multi-layer secured big data framework may serve to avoid potential network attacks. Anonymity algorithms may help to hide the identity of individual users during the use of their electricity consumption data.

- Higher-dimensional, multi-source industry data is another great challenge that needs to be considered. Cutting-edge advancement of deep learning neural model offers dimensionality reduction method of non-linear data *e.g.,* autoencoder. The autoencoders are unsupervised neural networks and can discover features with reduced dimensions. The autoencoders can be trained to use for forecasting as well. Moreover, the correlation of data can be explored by analysis of the decision rule conditions and its results, measuring relevance score, layer-wise propagation, etc. Therefore, to implement ML algorithms on big data, existing big data frameworks must be developed with a supportive processing system so that algorithms can be trained on the huge amount of big data and learn the pattern of data.
- The scarcity of datasets limits many efforts at big data innovation. Big data analytics depends on large-good quality datasets and sometimes labeled datasets. To predict or forecast events, datasets greatly assist data analytics methods. Data mining, ML algorithms are capable of real-time monitoring of seismic events from seismic data. We have to build large open-source datasets for mineral data analytics research to explore more. Through drilling simulation, virtual and physical data and parameters can be compared and uncertainty can be determined [43]. As a result, cost optimization is possible. Therefore, a data-driven simulator should be developed.
- Along with industry data, social media reviews are the primary source of consumer data and play a vital role in user and business analytics. Using advanced big data frameworks, users' experiences, comments, etc. can be collected and ML algorithms can efficiently discover the assessment of a product by users. The manufacturing industry should draw critical attention to consumers' review and collection of their demand, which may help to increase successful production growth.
- Finally, it is a very urgent need to build an open-source database of industry data. Then, the integration of ML algorithms with big data technologies can provide efficient and dynamic means of capturing latent features of reservoir parameters and data and correlation between them. Advanced ML techniques such as PCA, latent dirichlet allocation (LDA), various types of autoencoder, etc. offer not only feature extraction and covariance but also dimensionality reduction and data labeling facilities. Explainable AI tools e.g., local interpretable model-agnostic explanations (LIME), Shapley Additive Explanations (SHAP), facets, what-if tools, etc. can help to discover the explanation of the prediction properly and assists the process of model optimization.

VII. CONCLUDING REMARKS

This paper presented a review of papers related to big data technologies implemented in the power, mineral, and manufacturing industries. We also discussed the paper collection, selection, and assessment criteria done before the review task. For high-quality paper collection, we searched in IEEE digital library, ACM digital library, SpringerLink, Elsevier, Multidisciplinary Digital Publishing Institute (MDPI), Google Scholar, Willey, etc. and based on our selected criteria, we filtered only good quality papers for review. We proposed a taxonomy of applications of big data technologies in power, mineral, and manufacturing industries that have been demonstrated on three levels. Along with this, we presented the year-wise distribution and frequency of reviewed papers, and big data technologies used to acquire and process the massive amount of industry data. We also demonstrated the frequency of ML and data mining techniques that are used for industry data processing. Then, we discussed state-of-the-art big data technologies that have been proposed to collect, store, manage, and analyze power, mineral, and manufacturing data.

We have investigated the existing big data research gaps in the industry sectors and tried to bridge the gaps by recommending suggestions for data-driven industry approaches. The big data quality assessment framework need to be precisely upgraded for multi-dimensional big data to get outperforms the later processing and storage in the general industry. Moreover, noise cancellation big data framework may help to avoid noise during the transmitted data collection process which further improves the quality of big data. Besides, multi-layer secure protection is required for customers and business data for business data analysis. While for industry data analysis purposes, ML and data mining techniques proved tremendous efficiency, these tasks demand high-performance processing systems; therefore, the existing big data frameworks need to be re-framed to support ML and data mining algorithms/ models to be trained with big data.

For industry automation, big data can play a vital role by training intelligent agents and thus reduces human effort, cost, and production time. Introducing intelligent agents has crucial significance, especially in certain industrial environments to reduce life risk.

REFERENCES

[1] A. Tsanousa, E. Bektsis, C. Kyriakopoulos, A. G. González, U. Leturiondo, I. Gialampoukidis, A. Karakostas, S. Vrochidis, and I. Kompatsiaris, "A review of multisensor data fusion solutions in smart manufacturing: Systems and trends," *Sensors*, vol. 22, no. 5, p. 1734, Feb. 2022.

[2] H. Liu, F. Huang, H. Li, W. Liu, and T. Wang, "A big data framework for electric power data quality assessment," in *Proc. 14th Web Inf. Syst. Appl. Conf. (WISA)*, Nov. 2017, pp. 289–292.

[3] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid—A review," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 1099–1107, Nov. 2017.

[4] H. Patel, D. Prajapati, D. Mahida, and M. Shah, "Transforming petroleum downstream sector through big data: A holistic review," *J. Petroleum Explor. Prod. Technol.*, vol. 10, no. 6, pp. 2601–2611, Aug. 2020.

[5] M. Mohammadpoor and F. Torabi, "Big data analytics in oil and gas industry: An emerging trend," *Petroleum*, vol. 6, no. 4, pp. 321–328, Dec. 2020.

[6] P. Wu and J. Tan, "The design of distributed power big data analysis framework and its application in residential electricity analysis," in *Proc. 6th Int. Conf. Adv. Cloud Big Data (CBD)*, Aug. 2018, pp. 77–82.

[7] M. R. Islam, S. Sarker, M. S. Mazumder, and M. R. Ranim, "An IoT based real-time low cost smart energy meter monitoring system using Android application," 2020, *arXiv:2001.10350*.

[8] J. N. Desai, S. Pandian, and R. K. Vij, "Big data analytics in upstream oil and gas industries for sustainable exploration and development: A review," *Environ. Technol. Innov.*, vol. 21, Feb. 2021, Art. no. 101186.
[9] F. Murray, "Data challenges and opportunities for environmental management of north sea oil and gas decommissioning in an era of blue growth," *Mar. Policy*, vol. 97, pp. 130–138, Nov. 2018.
[10] W. Wu, X. Lu, B. Cox, G. Li, L. Lin, and Q. Yang, "Retrieving information and discovering knowledge from unstructured data using big data mining technique: Heavy oil fields example," in *Proc. Int. Petroleum Technol. Conf.*, Dec. 2014.
[11] T. Olneva, D. Kuzmin, S. Rasskazova, and A. Timirgalin, "Big data approach for geological study of the big region West Siberia," in *Proc. SPE Annu. Tech. Conf. Exhib.*, 2018.
[12] M. Akoum and A. Mahjoub, "A unified framework for implementing business intelligence, real-time operational intelligence and big data analytics for upstream oil industry operators," in *Proc. SPE Middle East Intell. Energy Conf. Exhib.*, 2013.
[13] W. Wang, "A model to determine the optimal critical level and the monitoring intervals in condition-based maintenance," *Int. J. Prod. Res.*, vol. 38, no. 6, pp. 1425–1436, 2000.
[14] R. Ak, O. Fink, and E. Zio, "Two machine learning approaches for short-term wind speed time-series prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1734–1747, Aug. 2016.
[15] C. Adrian, R. Abdullah, R. Atan, and Y. Y. Jusoh, "Towards developing strategic assessment model for big data implementation: A systematic literature review," *Int. J. Adv. Soft Comput. Appl.*, vol. 8, no. 3, pp. 1–20, 2016.
[16] M. R. Alam, K. M. Muttaqi, and A. Bouzerdoum, "Evaluating the effectiveness of a machine learning approach based on response time and reliability for islanding detection of distributed generation," *IET Renew. Power Gener.*, vol. 11, no. 11, pp. 1392–1400, Sep. 2017.
[17] L. Cadei, M. Montini, F. Landi, F. Porcelli, V. Michetti, M. Origgi, M. Tonegutti, and S. Duranton, "Big data advanced analytics to forecast operational upsets in upstream production system," in *Proc. Abu Dhabi Int. Petroleum Exhib. Conf.*, 2018.
[18] L. Wang, M. G. Mehrabi, and E. Kannatey-Asibu, "Hidden Markov model-based tool wear monitoring in turning," *J. Manuf. Sci. Eng.*, vol. 124, no. 3, pp. 651–658, Aug. 2002.
[19] A. Alfaleh, Y. Wang, B. Yan, J. Killough, H. Song, and C. Wei, "Topological data analysis to solve big data problem in reservoir engineering: Application to inverted 4D seismic data," in *Proc. SPE Annu. Tech. Conf. Exhib.*, Sep. 2015.
[20] R. M. Alguliyev, R. M. Aliguliyev, and M. S. Hajirahimova, "Big data integration architectural concepts for oil and gas industry," in *Proc. IEEE 10th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2016, pp. 1–5.
[21] B. T. Rollins, A. Broussard, B. Cummins, A. Smiley, and N. Dobbs, "Continental production allocation and analysis through big data," in *Proc. SPE/AAPG/SEG Unconventional Resour. Technol. Conf.*, 2017.
[22] S. L. Chan, Y. Lu, and Y. Wang, "Data-driven cost estimation for additive manufacturing in cybermanufacturing," *J. Manuf. Syst.*, vol. 46, pp. 115–126, Jan. 2018.
[23] T. Kourti and J. MacGregor, "Multivariate SPC methods for monitoring and diagnosing of process performance," in *Proc. PSE*, 1994, pp. 739–746.
[24] W. H. Allen, A. Rubaai, and R. Chawla, "Fuzzy neural network-based health monitoring for HVAC system variable-air-volume unit," *IEEE Trans. Ind. Appl.*, vol. 52, no. 3, pp. 2513–2524, May 2016.
[25] A. M. Crespino, A. Corallo, M. Lazoi, D. Barbagallo, A. Appice, and D. Malerba, "Anomaly detection in aerospace product manufacturing: Initial remarks," in *Proc. IEEE 2nd Int. Forum Res. Technol. Soc. Ind. Leveraging Better Tomorrow (RTSI)*, Sep. 2016, pp. 1–4.
[26] Z. Shah, A. Anwar, A. N. Mahmood, Z. Tari, and A. Y. Zomaya, "A spatiotemporal data summarization approach for real-time operation of smart grid," *IEEE Trans. Big Data*, vol. 6, no. 4, pp. 624–637, Dec. 2020.
[27] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, "Data and knowledge mining with big data towards smart production," *J. Ind. Inf. Integr.*, vol. 9, pp. 1–13, Mar. 2018.
[28] F. Yacoub and J. F. MacGregor, "Product optimization and control in the latent variable space of nonlinear PLS models," *Chemometrics Intell. Lab. Syst.*, vol. 70, no. 1, pp. 63–74, 2004.
[29] A. Baaziz and L. Quoniam, "How to use big data technologies to optimize operations in upstream petroleum industry," *Int. J. Innov.*, vol. 1, no. 1, pp. 19–25, Dec. 2013.
[30] R. Roden, "Seismic interpretation in the age of big data," in *Proc. SEG Int. Expo. Annu. Meeting*, 2016.
[31] E. Balouji and O. Salor, "Classification of power quality events using deep learning on event images," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 216–221.
[32] C. Fan, F. Xiao, Z. Li, and J. Wang, "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review," *Energy Buildings*, vol. 159, pp. 296–308, Jan. 2018.
[33] D. Li, Y. Gong, G. Tang, and Q. Huang, "Research and design of mineral resource management system based on big data and GIS technology," in *Proc. 5th IEEE Int. Conf. Big Data Anal. (ICBDA)*, May 2020, pp. 52–56.
[34] J. Betz, "Low oil prices increase value of big data in fracturing," *J. Petroleum Technol.*, vol. 67, no. 4, pp. 60–61, Apr. 2015.
[35] Y. Ren and Y. Ren, "A framework of data mining for logging reservoir evaluation," in *Proc. 13th Int. Conf. Service Syst. Service Manag. (ICSSSM)*, Jun. 2016, pp. 1–6.
[36] K. Bauman, A. Tuzhilin, and R. Zaczynski, "Using social sensors for detecting emergency events: A case of power outages in the electrical utility industry," *ACM Trans. Manag. Inf. Syst.*, vol. 8, nos. 2–3, pp. 1–20, Sep. 2017.
[37] 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.
[38] K. Govindan, T. C. E. Cheng, N. Mishra, and N. Shukla, "Big data analytics and application for logistics and supply chain management," *Transp. Res. E, Logistics Transp. Rev.*, vol. 114, pp. 343–349, Jun. 2018.
[39] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, May 2015.
[40] S. Gupta, L. Saputelli, and M. Nikolaou, "Big data analytics workflow to safeguard ESP operations in real-time," in *Proc. SPE North Amer. Artif. Lift Conf. Exhib.*, 2016.
[41] S. Garcia-Muñoz, T. Kourti, J. F. MacGregor, A. G. Mateos, and G. Murphy, "Troubleshooting of an industrial batch process using multivariate methods," *Ind. Eng. Chem. Res.*, vol. 42, no. 15, pp. 3592–3601, Jul. 2003.
[42] A. Sukapradja, J. Clark, H. Hermawan, and S. Tjiptowiyono, "Sisi Nubi dashboard: Implementation of business intelligence in reservoir modelling & synthesis: Managing big data and streamline the decision making process," in *Proc. SPE/IATMI Asia Pacific Oil Gas Conf. Exhib.*, 2017.
[43] M. Hutchinson, B. Thornton, P. Theys, and H. Bolt, "Optimizing drilling by simulation and automation with big data," in *Proc. SPE Annu. Tech. Conf. Exhib.*, Sep. 2018.
[44] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security Privacy*, vol. 7, no. 3, pp. 75–77, Jun. 2009.
[45] E. De Santis, L. Livi, A. Sadeghian, and A. Rizzi, "Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification," *Neurocomputing*, vol. 170, pp. 368–383, Dec. 2015.
[46] T. Li, Y. He, and C. Zhu, "Big data oriented macro-quality index based on customer satisfaction index and PLS-SEM for manufacturing industry," in *Proc. Int. Conf. Ind. Informat.-Comput. Technol., Intell. Technol., Ind. Inf. Integr. (ICIICII)*, Dec. 2016, pp. 181–186.
[47] O. Bello, D. Yang, S. Lazarus, X. S. Wang, and T. Denney, "Next generation downhole big data platform for dynamic data-driven well and reservoir management," in *Proc. SPE Reservoir Characterisation Simul. Conf. Exhib.*, May 2017.
[48] P. Joshi, R. Thapliyal, A. Chittambakkam, R. Ghosh, S. Bhowmick, and S. Khan, "Big data analytics for micro-seismic monitoring," in *Proc. Offshore Technol. Conf. Asia*, 2018.
[49] C. M. Jaeckle and J. F. MacGregor, "Industrial applications of product design through the inversion of latent variable models," *Chemometric Intell. Lab. Syst.*, vol. 50, no. 2, pp. 199–210, Mar. 2000.
[50] F. Han and S. Z. Sun, "Petroleum geoscience big data and GPU parallel computing," in *Proc. IEEE Int. Conf. Multimedia Big Data*, Apr. 2015, pp. 292–293.
[51] X. Li, J. Song, and B. Huang, "A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics," *Int. J. Adv. Manuf. Technol.*, vol. 84, nos. 1–4, pp. 119–131, Apr. 2016.
[52] S. Baek and D.-Y. Kim, "Abrupt variance and discernibility analyses of multi-sensor signals for fault pattern extraction," *Comput. Ind. Eng.*, vol. 128, pp. 999–1007, Feb. 2019.
[53] J. Flores-Cerrillo and J. F. MacGregor, "Control of batch product quality by trajectory manipulation using latent variable models," *J. Process Control*, vol. 14, no. 5, pp. 539–553, Aug. 2004.

[54] A. B. Mahfoodh, M. Ibrahim, M. Hawi, and K. Hakami, "Introducing a big data system for maintaining well data quality and integrity in a world of heterogeneous environment," in *Proc. SPE Kingdom Saudi Arabia Annu. Tech. Symp. Exhib.*, 2017.

[55] D. Khvostichenko and S. Makarychev-Mikhailov, "Effect of fracturing chemicals on well productivity: Avoiding pitfalls in big data analysis," in *Proc. SPE Int. Conf. Exhib. Formation Damage Control*, 2018.

[56] P. Baruah and R. B. Chinnam, "HMMs for diagnostics and prognostics in machining processes," *Int. J. Prod. Res.*, vol. 43, no. 6, pp. 1275–1293, Mar. 2005.

[57] M. K. Jena and S. R. Samantaray, "Data-mining-based intelligent differential relaying for transmission lines including UPFC and wind farms," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 8–17, Jan. 2016.

[58] D. Liu, L. Zeng, C. Li, K. Ma, Y. Chen, and Y. Cao, "A distributed short-term load forecasting method based on local weather information," *IEEE Syst. J.*, vol. 12, no. 1, pp. 208–215, Mar. 2018.

[59] E. De Santis, A. Rizzi, and A. Sadeghian, "A learning intelligent system for classification and characterization of localized faults in smart grids," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2017, pp. 2669–2676.

[60] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: A review," *Energy Informat.*, vol. 1, no. 1, pp. 1–24, Dec. 2018.

[61] N. K. Dev, R. Shankar, R. Gupta, and J. Dong, "Multi-criteria evaluation of real-time key performance indicators of supply chain with consideration of big data architecture," *Comput. Ind. Eng.*, vol. 128, pp. 1076–1087, Feb. 2019.

[62] M. Tanabe and A. Miyake, "Safety design approach for onshore modularized LNG liquefaction plant," *J. Loss Prevention Process Industries*, vol. 23, no. 4, pp. 507–514, Jul. 2010.

[63] Y. Tang, C. W. Ten, C. Wang, and G. Parker, "Extraction of energy information from analog meters using image processing," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2032–2040, Jul. 2015.

[64] Z. Liu, Y. Wang, L. Cai, Q. Cheng, and H. Zhang, "Design and manufacturing model of customized hydrostatic bearing system based on cloud and big data technology," *Int. J. Adv. Manuf. Technol.*, vol. 84, nos. 1–4, pp. 261–273, Apr. 2016.

[65] W. Duffy, J. Rigg, and E. Maidla, "Efficiency improvement in the Bakken realized through drilling data processing automation and the recognition and standardization of best safe practices," in *Proc. SPE/IADC Drilling Conf. Exhib.*, 2017.

[66] D. Dinis, A. Barbosa-Povoa, and A. P. Teixeira, "Valuing data in aircraft maintenance through big data analytics: A probabilistic approach for capacity planning using Bayesian networks," *Comput. Ind. Eng.*, vol. 128, pp. 920–936, Feb. 2019.

[67] L. Canetta, A. Salvade, P. Schnegg, E. Müller, and M. Lanini, "RFID–ERP key data integration challenges," in *Digital Factory for Human-Oriented Production Systems*. Berlin, Germany: Springer, 2011, pp. 73–95.

[68] J. Johnston and A. Guichard, "New findings in drilling and wells using big data analytics," in *Proc. Offshore Technol. Conf.*, 2015.

[69] M. Nino, F. Saenz, J. M. Blanco, and A. Illarramendi, "Requirements for a big data capturing and integration architecture in a distributed manufacturing scenario," in *Proc. IEEE 14th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2016, pp. 1326–1329.

[70] X. Han, X. Wang, and H. Fan, "Requirements analysis and application research of big data in power network dispatching and planning," in *Proc. IEEE 3rd Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Oct. 2017, pp. 663–668.

[71] L. Wei, Q. Zhao, and H. Shu, "Design of manufacturing big data access platform based on SOA," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 1841–1845.

[72] A. Kumar, R. B. Chinnam, and F. Tseng, "An HMM and polynomial regression based approach for remaining useful life and health state estimation of cutting tools," *Comput. Ind. Eng.*, vol. 128, pp. 1008–1014, Feb. 2019.

[73] C. Chelmis, J. Zhao, V. Sorathia, S. Agarwal, and V. Prasanna, "Semiautomatic, semantic assistance to manual curation of data in smart oil fields," in *Proc. SPE Western Regional Meeting*, 2012.

[74] S. Kar, S. R. Samantaray, and M. D. Zadeh, "Data-mining model based intelligent differential microgrid protection scheme," *IEEE Syst. J.*, vol. 11, no. 2, pp. 1161–1169, Jun. 2017.

[75] T. Palmer and M. Turland, "Proactive rod pump optimization: Leveraging big data to accelerate and improve operations," in *Proc. SPE North Amer. Artif. Lift Conf. Exhib.*, 2016.

[76] F. Hashemi, M. Mohammadi, and A. Kargarian, "Islanding detection method for microgrid based on extracted features from differential transient rate of change of frequency," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 4, pp. 891–904, Mar. 2017.

[77] N. Rossi, J. Michelez, and F. Concina, "Technology update: Big data for advanced well engineering holds strong potential to optimize drilling costs," *J. Petroleum Technol.*, vol. 70, no. 2, pp. 18–21, Feb. 2018.

[78] M. Song and X. Zhou, "A casing damage prediction method based on principal component analysis and gradient boosting decision tree algorithm," in *Proc. SPE Middle East Oil Gas Show Conf.*, 2019.

[79] Z. Fan, P. Kulkarni, S. Gormus, C. Efthymiou, G. Kalogridis, M. Sooriyabandara, Z. Zhu, S. Lambotharan, and W. H. Chin, "Smart grid communications: Overview of research challenges, solutions, and standardization activities," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 21–38, 1st Quart., 2013.

[80] N. Li, M. Xu, W. Cao, and P. Gao, "Researches on data processing and data preventing technologies in the environment of big data in power system," in *Proc. 5th Int. Conf. Electric Utility Deregulation Restructuring Power Technol. (DRPT)*, Nov. 2015, pp. 2491–2494.

[81] G. Sheng, H. Hou, X. Jiang, and Y. Chen, "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 695–702, Mar. 2018.

[82] M. Layouni, M. S. Hamdi, and S. Tahar, "Detection and sizing of metal-loss defects in oil and gas pipelines using pattern-adapted wavelets and machine learning," *Appl. Soft Comput.*, vol. 52, pp. 247–261, Mar. 2017.

[83] E. Maidla, W. Maidla, J. Rigg, M. Crumrine, and P. Wolf-Zoellner, "Drilling analysis using big data has been misused and abused," in *Proc. IADC/SPE Drilling Conf. Exhib.*, 2018.

[84] Z. Lv, W. Deng, Z. Zhang, N. Guo, and G. Yan, "A data fusion and data cleaning system for smart grids big data," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Dec. 2019, pp. 802–807.

[85] S. Li, P. Wang, and L. Goel, "A novel wavelet-based ensemble method for short-term load forecasting with hybrid neural networks and feature selection," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1788–1798, May 2016.

[86] H. Sun, Z. Wang, and J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2516–2524, Sep. 2016.

[87] V. Malbasa, C. Zheng, P.-C. Chen, T. Popovic, and M. Kezunovic, "Voltage stability prediction using active machine learning," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 3117–3124, Nov. 2017.

[88] N. Shukla, M. K. Tiwari, and G. Beydoun, "Next generation smart manufacturing and service systems using big data analytics," *Comput. Ind. Eng.*, vol. 128, pp. 905–910, Feb. 2019.

[89] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, Oct. 2015.

[90] W. Yuanting, Z. Zhanmin, D. Lihong, X. Weiyi, M. Guizhi, Z. Dengwen, L. Yun, Y. Bo, W. Jiaheng, and Y. Limin, "Research and application of big data analysis platform for oil production engineering in huabei oilfield," in *Proc. IEEE 4th Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2019, pp. 148–151.

[91] G. D. F. Morales, "SAMOA: A platform for mining big data streams," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 777–778.

[92] A. Swetapadma and A. Yadav, "Data-mining-based fault during power swing identification in power transmission system," *IET Sci. Meas. Technol.*, vol. 10, no. 2, pp. 130–139, Mar. 2016.

[93] P. N. Papadopoulos, T. Guo, and J. V. Milanovic, "Probabilistic framework for online identification of dynamic behavior of power systems with renewable generation," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 45–54, Jan. 2018.

[94] B. Huang, C. Li, C. Yin, and X. Zhao, "Cloud manufacturing service platform for small- and medium-sized enterprises," *Int. J. Adv. Manuf. Technol.*, vol. 65, nos. 9–12, pp. 1261–1272, Apr. 2013.

[95] J. F. MacGregor, M. Bruwer, I. Miletic, M. Cardin, and Z. Liu, "Latent variable models and big data in the process industries," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 520–524, 2015.

[96] J. Wang, X. Xiong, N. Zhou, Z. Li, and W. Wang, "Early warning method for transmission line galloping based on SVM and AdaBoost Bi-level classifiers," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 14, pp. 3499–3507, Nov. 2016.

[97] N. P. Sarapulov and R. A. Khabibullin, "Application of big data tools for unstructured data analysis to improve ESP operation efficiency," in *Proc. SPE Russian Petroleum Technol. Conf.*, Oct. 2017.
[98] M. Ockree, K. G. Brown, J. Frantz, M. Deasy, and R. John, "Integrating big data analytics into development planning optimization," in *Proc. SPE/AAPG Eastern Regional Meeting*, 2018.
[99] Y. Zhang, P. Peng, C. Liu, and H. Zhang, "Anomaly detection for industry product quality inspection based on Gaussian restricted Boltzmann machine," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 1–6.
[100] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *Proc. IREP Symp. Bulk Power Syst. Dyn. Control Optim., Secur. Control Emerg. Power Grid*, Aug. 2013, pp. 1–9.
[101] D. P. Mishra, S. R. Samantaray, and G. Joos, "A combined wavelet and data-mining based intelligent protection scheme for microgrid," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2295–2304, Sep. 2016.
[102] 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.
[103] X. Wang, S. D. J. McArthur, S. M. Strachan, J. D. Kirkwood, and B. Paisley, "A data analytic approach to automatic fault diagnosis and prognosis for distribution automation," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6265–6273, Nov. 2018.
[104] C. L. Stergiou and K. E. Psannis, "Digital twin intelligent system for industrial IoT-based big data management and analysis in cloud," *Virtual Reality Intell. Hardw.*, vol. 4, no. 4, pp. 279–291, Aug. 2022.
[105] D. Seemann, M. Williamson, and S. Hasan, "Improving resevoir management through big data technologies," in *Proc. SPE Middle East Intell. Energy Conf. Exhib.*, 2013.
[106] M. Yang, Y. Lin, and X. Han, "Probabilistic wind generation forecast based on sparse Bayesian classification and Dempster–Shafer theory," *IEEE Trans. Ind. Appl.*, vol. 52, no. 3, pp. 1998–2005, Jun. 2016.
[107] L. Zhu, C. Lu, Z. Y. Dong, and C. Hong, "Imbalance learning machine-based power system short-term voltage stability assessment," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2533–2543, Oct. 2017.
[108] A. Singh, N. Shukla, and N. Mishra, "Social media data analytics to improve supply chain management in food industries," *Transp. Res. E, Logistics Transp. Rev.*, vol. 114, pp. 398–415, Jun. 2018.
[109] S.-B. Huang, X.-X. Yang, L.-S. Shen, and Y.-M. Li, "Fuzzy co-clustering algorithm for high-order heterogeneous data," *J. Commun.*, vol. 35, no. 6, p. 15, 2014.
[110] P. O'Donovan, K. Leahy, K. Bruton, and D. T. J. O'Sullivan, "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities," *J. Big Data*, vol. 2, no. 1, pp. 1–26, Dec. 2015.
[111] D. Zhang, S. Li, M. Sun, and Z. O'Neill, "An optimal and learning-based demand response and home energy management system," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1790–1801, Jul. 2016.
[112] Y. Zhang, Y. Xu, Z. Y. Dong, Z. Xu, and K. P. Wong, "Intelligent early warning of power system dynamic insecurity risk: Toward optimal accuracy-earliness tradeoff," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2544–2554, Oct. 2017.
[113] I. Taleb, M. A. Serhani, and R. Dssouli, "Big data quality assessment model for unstructured data," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2018, pp. 69–74.
[114] A. Lin, "Principles of big data algorithms and application for unconventional oil and gas resources," in *Proc. SPE Large Scale Comput. Big Data Challenges Reservoir Simul. Conf. Exhib.*, 2014.
[115] S. Raphael, C. P. Fuge, S. Gutierrez, H. A. Kuzma, and N. S. Arora, "Big data every day: Predictive analytics used to improve production surveillance," in *Proc. SPE Digit. Energy Conf. Exhib.*, 2015.
[116] J. Siryani, B. Tanju, and T. J. Eveleigh, "A machine learning decision-support system improves the Internet of Things smart meter operations," *IEEE Internet Things J.*, vol. 4, no. 4, pp. 1056–1066, Aug. 2017.
[117] M. Yuan, K. Liu, L. Zhang, and C. Zou, "Research on big data storage model of oilfield assay data based on MongoDB," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 1863–1866.
[118] Y. Ren, P. N. Suganthan, and N. Srikanth, "A novel empirical mode decomposition with support vector regression for wind speed forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1793–1798, Aug. 2016.
[119] C. Sousa, I. H. F. Santos, V. T. Almeida, A. R. Almeida, G. M. Silva, A. E. Ciarlini, A. Prado, R. D. A. Senra, V. Gottin, A. Bhaya, and T. Calmon, "Applying big data analytics to logistics processes of oil and gas exploration and production through a hybrid modeling and simulation approach," in *Proc. OTC Brasil*, 2015.
[120] C. Zhang and F. Wang, "Multi-feature fusion based anomaly electro-data detection in smart grid," in *Proc. 15th Int. Symp. Pervasive Syst., Algorithms Netw. (I-SPAN)*, Oct. 2018, pp. 54–59.
[121] F. Tao, Y. Cheng, L. D. Xu, L. Zhang, and B. H. Li, "CCIoT-CMfg: Cloud computing and Internet of Things-based cloud manufacturing service system," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1435–1442, May 2014.
[122] C. Zhao, Z. Wu, and X. Jiang, "Cross-platform data fusion and storage pattern of power grid planning," *Electr. Power Construct.*, vol. 36, no. 3, pp. 119–122, 2015.
[123] E. Udegbe, E. Morgan, and S. Srinivasan, "From face detection to fractured reservoir characterization: Big data analytics for restimulation candidate selection," in *Proc. SPE Annu. Tech. Conf. Exhib.*, 2017.
[124] W. Wang, D. Forrester, and P. Frith, "A generalized machine fault detection method using unified change detection," Defence Sci. Technol. Organisation, Melbourne, U.K., Tech. Rep., 2014.
[125] E. Vezzetti, M. Alemanni, C. Balbo, and A. L. Guerra, "Big data analysis techniques for supporting product lifecycle management in the fashion industries," in *Proc. Workshop Bus. Models ICT Technol. Fashion Supply Chain*. Cham, Switzerland: Springer, 2017, pp. 25–34.
[126] Z. Wang, B. Wu, D. Bai, and J. Qin, "Distributed big data mining platform for smart grid," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 2345–2354.
[127] M. Zhang, Z. Ye, and M. Xie, "A condition-based maintenance strategy for heterogeneous populations," *Comput. Ind. Eng.*, vol. 77, pp. 103–114, Nov. 2014.
[128] W. Wu and M. Peng, "A data mining approach combining K-means clustering with bagging neural network for short-term wind power forecasting," *IEEE Internet Things J.*, vol. 4, no. 4, pp. 979–986, Aug. 2017.
[129] T. Zhao, Y. Zhang, and D. Zhang, "Application technology of big data in smart distribution grid and its prospect analysis," *Power Syst. Technol.*, vol. 38, no. 12, pp. 3305–3312, 2014.
[130] L. Xianglan, "Digital construction of coal mine big data for different platforms based on life cycle," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 456–459.
[131] J. Xiao and X. Sun, "Big data analytics drive EOR projects," in *Proc. SPE Offshore Eur. Conf. Exhib.*, Sep. 2017.
[132] B. Yao, W. Bin, C. Yufeng, and B. Demeng, "BDAP: A data mining platform based on spark," *J. Univ. Sci. Technol. China*, vol. 47, no. 4, p. 358, 2017.
[133] NTNU. (2021). *Norne Model*. Accessed: May 26, 2022. [Online]. Available: http://www.ipt.ntnu.no/~norne/wiki1/doku.php?id=english:nornebenchmarkcase2
[134] (2021). *Canadian Energy Pipeline Association*. Accessed: Apr. 4, 2022. [Online]. Available: https://www.cepa.com/library/factoids
[135] Q. Liu, C. Chen, and S. Chen, "Key technology of intelligentized welding manufacturing and systems based on the Internet of Things and multi-agent," *J. Manuf. Mater. Process.*, vol. 6, no. 6, p. 135, Nov. 2022.
[136] Y. Jiang, W. Wang, and C. Zhao, "A machine vision-based realtime anomaly detection method for industrial products using deep learning," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 4842–4847.

SUPRIYA SARKER received the B.Sc. and M.Sc. degrees in computer science and engineering from the Chittagong University of Engineering and Technology (CUET), Chattogram, Bangladesh, in 2014 and 2022, respectively. She is currently pursuing the Doctoral degree with the University of Memphis, Memphis, TN, USA. She is also working as a Graduate Assistant at the Department of Computer Science, University of Memphis. Her research interests include intelligent transportation systems, autonomous vehicles, machine learning, data science, and big data. She regularly serves as a Reviewer for IEEE conferences and reputed journals, such as IEEE Access, Elsevier, and Springer.

MOHAMMAD SHAMSUL AREFIN (Senior Member, IEEE) received the Doctor of Engineering degree in information engineering from Hiroshima University, Japan, with the support of the scholarship of MEXT, Japan. As a part of his doctoral research, he was with IBM Yamato Software Laboratory, Japan. He is in lien at the Chittagong University of Engineering and Technology (CUET), Chattogram, Bangladesh, and currently affiliated with the Department of Computer Science and Engineering (CSE), Daffodil International University, Dhaka, Bangladesh. Earlier, he was the Head of the Department of Computer Science and Engineering, CUET. He has more than 110 refereed publications in international journals, book series, and conference proceedings. His research interests include privacy-preserving data publishing and mining, distributed and cloud computing, big data management, multilingual data management, semantic web, object-oriented system development, and IT for agriculture and the environment. He is a member of ACM and a fellow of IEB and BCS. He is the Organizing Chair of BIM 2021, the TPC Chair of ECCE 2017, the Organizing Co-Chair of ECCE 2019, and the Organizing Chair of BDML 2020. He has visited Japan, Indonesia, Malaysia, Bhutan, Singapore, South Korea, Egypt, India, Saudi Arabia, and China, for different professional and social activities.

MD KOWSHER is currently pursuing the Ph.D. degree with the Stevens Institute of Technology. In previous, he worked as an Artificial Intelligence Scientist at Hishab Ltd., and an AI Engineer at NKSoft, USA. He is currently working as a Doctoral Research Assistant at Stevens AI Laboratory, USA. He received the Provost Fellowship Award. He also received the Best Paper Awards from various international conferences, such as ICONCS, IC4ME2, ICCCM, NISS, and ICIET. Apart from that, he was the Champion of Robi r-venture 2.0 and received the National Basis ICT Award in 2019. In 2021, he received the Scientist of the Year Award, for his excellent research in the field of AI from IEM, India. He reviewed many papers in ICCIDM, ICSECS, ICOCSIM, and *Visual Computing for Industry, Biomedicine, and Art*.

TOUHID BHUIYAN received the Ph.D. degree in information security, focusing on trust management for intelligent recommendations from the Queensland University of Technology (QUT), Australia. He is currently the Head of the Department of Computer Science and Engineering, Daffodil International University (DIU), Bangladesh. He is a Certified Ethical Hacker. He was the Director of the Cyber Security Centre, DIU, where he was also the Head and a Professor of the Software Engineering Department. Before joining at AIB, he worked for several institutions, including Monash University, The University of Western Australia, QUT, University of Western Sydney, and Central College Sydney. He has more than 114 research publications in renowned national and international journals, books, and conference proceedings. His research interests include cyber security, intelligent recommendations, social networks, trust management, big data analytics, e-health, e-learning, artificial intelligence, online learning, database management, and software engineering. He has received the Cyber Security: Cyber Risk and Resilience Certificate from the University of Oxford. He was a recipient of the Australian Postgraduate Award (APA) and the Deputy Vice-Chancellor's Initiative Scholarship from QUT.

PRANAB KUMAR DHAR received the B.Sc. degree from the Chittagong University of Engineering and Technology (CUET), Chattogram, Bangladesh, in 2004, the M.Sc. degree from the University of Ulsan, Republic of Korea, in 2010, and the Ph.D. degree from Saitama University, Japan, in 2014. In 2005, he joined as a Lecturer at the Department of Computer Science and Engineering, CUET, where he is currently serving as a Professor. He has published over 30 refereed journal articles and 40 conference papers. He is the author of two books, one book chapter, and one patent. His research interests include multimedia security, digital watermarking, steganography, multimedia data compression, sound synthesis, digital image processing, and digital signal processing. He is a member of the technical committee of several international conferences. He serves as a Reviewer for various reputed journals, including IEEE, IEICE, Elsevier, and Springer.

OH-JIN KWON received the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 1994. He was a Researcher with the Agency for Defense Development, South Korea, from 1984 to 1989, and the Head of the Media Laboratory, Samsung SDS Company Ltd., Seoul, South Korea, from 1995 to 1999. Since 1999, he has been a Faculty Member with Sejong University, Seoul, where he is currently a Professor. His research interests include image and video fusion, coding, watermarking, analysis, and processing.