Data science leverage and big data analysis for Internet of Things energy systems

6

Arman Behnam[1], *Sasan Azad*[2,3], *Mohammadreza Daneshvar*[4], *Amjad Anvari-Moghaddam*[5] *and Mousa Marzband*[6,7]

[1]Department of Computer Science, Illinois Institute of Technology, Chicago, United States, [2]Department of Electrical Engineering, Shahid Beheshti University, Tehran, Iran, [3]Electrical Networks Research Institute, Shahid Beheshti University, Tehran, Iran, [4]Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, [5]Department of Energy (AAU Energy), Aalborg University, Aalborg, Denmark, [6]Electrical Power and Control Systems Research Group, Northumbria University, Newcastle upon Tyne, United Kingdom, [7]Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah, Saudi Arabia

Chapter Outline

IoT Enabled Multi-Energy Systems. DOI: https://doi.org/10.1016/B978-0-323-95421-1.00006-9

6.1 Introduction

Smart grids (SGs) are the intelligent version of power grids well unified with connectivity and information infrastructures to make system monitoring, control, and grid management better. The two-way electricity flow and circulation of data in these structures are the most popular original attributes comparing traditional power grids. For conceiving a generic intuition about big data, the literature history, composition, methods, and facilities are needed to be discussed. The foundation of big data comes from the huge dataset processing issues and lacks suitable storage capabilities [1]. The big data evolution stages are categorized from megabyte (MB) to gigabyte (GB) from the 1970s to late 1980s, GB to terabyte (TB) from the late 1980s to late 1990, TB to petabyte (PB) from the late 1990s to 2012, and PB to exabyte announced commonly in 2012 [2]. There are many disputes about the exact and appropriate big data definition in vast approaches, and it sounds that attaining a consensus and institutionalized description has been always hard. Currently, roughly the popular three kinds of descriptions are architectural, attributive, and comparative explanations [3]. In addition, the term Internet of Things (IoT) itself is hard to define, since it produces great constant data quantity, which is related straight to the storage capacity. Big data is a term for massive structured, unstructured, and semistructured data, which has been used for information generation [4].

The sections of this chapter are presented as follows. In the second section, data science (DS) approaches and tools are discussed. In the third section, big data analytics methods are described. In the fourth section, future potential studies are being considered. In the fifth section, the conclusion of the chapter is brought.

6.2 Data science

With a large amount of structured data available in energy systems gathered by devices and sensors, there is a crucial need for new analytics methods for extracting knowledge out of these datasets [5]. In this section, the theoretical and practical structures of DS from SG problems to final products are discussed assisting the data scientists in solving real-world problems. Many keywords in this field, including data analysis, data mining (DM), big data, DS, machine learning (ML), and deep learning (DL), have been used among experts, which are highly correlated and challenging [6]. The "data analysis" is defined as the data processing by empirical or logical tools for knowledge extraction and practical purposes, while the "data analytics" is defined as instruments and processes that help with an in-depth information insight exploration [7]. Also, "data mining" is referred to knowledge discovery from data, data/pattern analysis, data archeology, and data dredging. In this regard, data sources consist of databases, data centers (DCs), big data warehouses, and the internet [8].

The term "big data" contains many more features and challenges, including massive, noisy, and erroneous data records, high dimensional data, heterogeneous

features, and unstructured data types [9]. In energy systems, big data is generated by IoT networks and devices. There are 5Vs, including volume, velocity, variety, veracity, and value, being used to understand and describe big data [10]. For data understanding purposes and granular data analysis, "advanced analytics" is defined as autonomous content analysis enabled by advanced techniques to discover deep insights and make recommendations in a new SG intelligence or analytics form [11]. Hence, the term "machine learning" is known as a branch of artificial intelligence (AI) with the goal of automating analytical model building. In addition, trends recognition and decision-making with minimal human involvement are recognized as ML purposes. The term "deep learning" is a subset in the ML field, inspired by the human brain's system and its operation termed artificial neural network (ANN) [12]. Hence, distinct from clear definitions discussed till now, the term "data science" is a conceptual field that comprises advanced data analytics, DM, ML, DL, modeling, and some other related disciplines such as statistics, optimization, ranking strategies, and useful information extraction and transforms them into SG decisions [13]. DS is introduced as a new interdisciplinary field that synthesizes statistics, information systems, optimization, communication knowledge, management, and hardware/software engineering to study data and its derivatives by employing a data-to-information-to-knowledge procedure [14]. In the following sections, all aspects of the DS comprehension and implementation are discussed.

6.2.1 Understanding data science modeling in smart grids

To understand how DS is able to play a vital role in real-world problems, different data structures should be categorized at first and then DS steps should be determined from SG understanding to final product and computerization. Commonly, data availability is the fundamental issue in data-driven application system building. The data is categorized in four fields: (1) structured—built up on a well-defined data structure following an approved order, that is, first and last names, phone numbers, addresses, credit/debit card information, stock market reports, and location; (2) unstructured—without any predefined format, that is, input data, e-mails, blog posts, text issues, text documents, audio, videos, images, presentations' files, and web pages; (3) semistructured—providing features from both the structured and unstructured data, that is, HTML, XML, JSON files, and NoSQL/MySQL databases; and (4) metadata—representing data as the data representator, that is, the author information, keywords, file type and size, file creation date, last modification date, and time [15]. DS model development, which is briefly explained in continue, helps scholars for data analysis in a discussed problem domain and gain insights out of the data to achieve a data-driven model or a product [16].

6.2.2 Steps of data science modeling in smart grid

DS model development from collected data to data product is shown in Fig. 6.1. In the following, each step of this DS process is concisely explained:

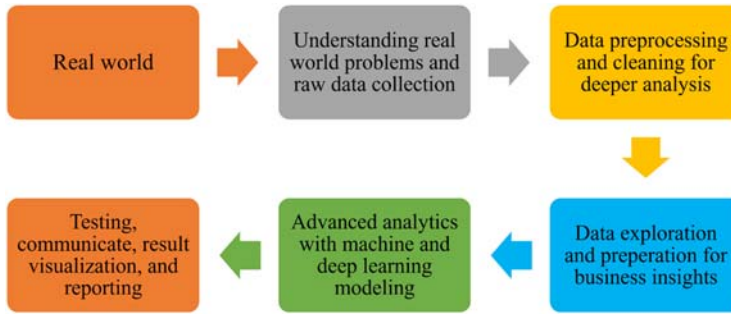

Figure 6.1 Data science modeling from real-world data to the data-driven system and decision-making.

- *The SG domain understanding*: Achieving a complete problem understanding throughout the SG domain is an initial task considering the impacts on the relevant firms/individuals and the ultimate targets. In this regard, there should be questions including which category/group is the target, which action/option should be chosen, etc. These questions depend on the problem's nature and are recommended to be asked at the beginning. These activities are essential to find more information about SG requirements and the expected information from data. This will result in enabling firms to improve their decision-making system, which is recognized as "SG Intelligence" [17]. Another vital step is the data source identification that will help with formulated question answering, also actions and trends based on the data. When the SG problem is defined clearly, the analytical tool or approach is used for the problem-solving.
- *Understanding data*: With a data-driven model/system, fine data understanding is the next step. The data preprocessing tasks are needed to achieve informative insights, which is crucial to any DS employment. Thus data evaluation that assesses the availability of data and its adjustment to the SG problem is known as the data understanding the first job. Several characteristics include data format, the quantity of dataset, quantity sufficiency, relevancy, data authorized access, variable importance, linking several data sources, data evaluation outputs, etc. [18]. They are undoubtedly essential for understanding the data for a discussed SG problem. Generally, the data understanding pipeline helps know the needs of data and the best ways to obtain it.
- *Data preprocessing*: In the beginning, a statistical developed model is needed besides providing tools for hypotheses creation by mostly data visualization and interpretation such as charts, pi-plots, and histograms. The data quality assurance, which is available by the data preprocessing methods, is generally the raw data transformation and cleaning procedure before data analytics. It is also associated with information format manipulating, adjusting data, and combining datasets to refine data. Then, several activities, including missing values imputation and manipulation, unbalanced data and bias problem-solving methods, statistical distribution modeling, dealing with outliers/anomalies, etc., are known as the key elements in this phase [19].
- *Machine learning methods and model evaluation*: Whenever the data is preprocessed for the model building, data scientists are ready to develop a model/algorithm to address the target problem. They are commonly distinct training and test sets of the given data. The division is mostly in the ratio of 80 to 20 or consideration of the most standard k-fold data partition tools [20]. To maximize the model performance, it should be observed every

moment. Several model validation and evaluation metrics, including accuracy, true and false positive/negative rates, precision, recall, F-score, error rate, (receiver operating characteristic curve) ROC, etc., can be used to evaluate the model performance, which helps to choose or design the learning structure [21]. Furthermore, many advanced analytics tools and approaches, including feature engineering/selection/extraction, parameter tuning, ensemble methods, etc., are available for machine learning experts to shape the final data-driven model to deal with a specific SG problem by smart decision-making.

- *Data science product*: Every DS activity output must be a DS product. A DS product generally is a data-enabled product, which is either SG exploration, prediction, data as a service, recommendation engine, data insight, making decisions, classification, clustering, knowledge, paradigm, application, or system. Thus, for decision-making improvement, great distinct machine learning pipelines and DS procedures should be developed [22].

Overall, implementation of DS methods can affect SG practices to alter the performance. The most appealing section in the DS procedure is attaining a further comprehension of the SG problem. If it does not happen, data collecting will be more difficult, which ends in performance decreasing for useful information extraction and decision-making. Talking about the role, "data scientists" usually interpret and study the data to expose the most substantial questions' answers, so organizations'/firms' decision-making procedure will be smoother [23]. In conclusion, a data scientist actively collects and interprets datasets out of many data sources, so a better understanding of the SG performance will be achieved. They also design and develop machine learning methods/algorithms, with attention to multivariational analytics, resulting in intelligent computing.

6.2.3 Advanced data analytics and smart computing in smart grids

Forecasting trends, episodes, and attitudes are cases used for advanced analytics methods and smart computing capabilities. Hence, advanced data analytics is known as automated or semiautomated content analysis to discover more informative insights, prove assumptions/hypotheses, and predict trends and make recommendations where machine learning comes to help [24]. In this regard, key SG questions should be asked, that is, "What is happening?," "Why did it happen?," "What will happen in the future?," and "What operation should be taken?." According to these questions, the data analytics will be divided into four categories, including descriptive, diagnostic, predictive, and prescriptive.

1. *Descriptive analytics*: Finding out the SG changes by historical data interpretation is categorized as descriptive analytics [25]. Thus, to answer the question "what has happened?," historical data summarization has been used. To present a precise SG problem's picture and its relation to previous times, descriptive analytics is important since utilizing a vast range of data is considered in this regard. At last, the definition of strength/weakness points is determined by administrators and managers.
2. *Diagnostic analytics*: The examination of content to answer the question, "Why did it happen?" [25]. Simply finding the problem origins is the goal of this type of analytics. It also enables value extraction from the dataset by suggesting the necessary questions and

deeply interpreting them to attain helpful answers. It is identified by techniques including data distribution interpretation, data visualization, and DM and recognizing the correlations.

3. *Predictive analytics*: It is an essential data analytics method for various purposes and several applications such as SG risk management, power consumption patterns, and predictive and prescriptive maintenance, enhancing their performance [26]. Modern SGs, for instance, simply use predictive analytics for cost minimization purposes by forecasting future demand trends and managing power flow and inventory, production capacity optimization, etc. As a result, predictive analytics is named as the DS analytical core.
4. *Prescriptive analytics*: For the overall rewards and profitability maximization purposes, prescriptive analytics finds the most usage out of available information, so the question "What operation/action should be taken?" can be answered [27]. Prescriptive analytics is known as the last part of the SG analytics, which gathers data from several predictive/descriptive data sources for the application of making decisions. So, its relation with predictive/descriptive analytics is obvious, but the difference is in the priority of actionable insights instead of data monitoring. Integrating ML, DL, SG domain knowledge, big data, and prescriptive analytics aids in the data-driven decision-making process for firms and organizations.

In conclusion, the reasons and the clues of occurrence should be clarified. Hence, descriptive and diagnostic analytics will operate historically. Historical records of data aid in predictive analytics and prescriptive analytics, so the steps that should be taken to make an impact on the discussed analytics will be clear. Forward-looking firms/organizations in the SG practices can use all these analytical methods for smart decision-making. All these results are summarized in Table. 6.1.

6.2.4 Supervised and unsupervised learning in smart grids

There are two learning algorithms that are fundamental categories of ML. While we learn about them early in our DS journey, we might not fully understand their differences, their use, and how we should approach them as engineering problems. Supervised learning is a machine learning method where output vectors and target

Table 6.1 Analytics methods and their description.

Analytical method types	Question answered by the method	Examples
Descriptive analytics	What happened in the past?	Summary of historical events
Diagnostic analytics	Why did it happen?	Anomaly detection and casual relations and effects
Predictive analytics	What is the result in the future?	Grid outputs prediction, recommendation engines, and availability and planning prediction
Prescriptive analytics	What operation should be taken?	SG management improvement

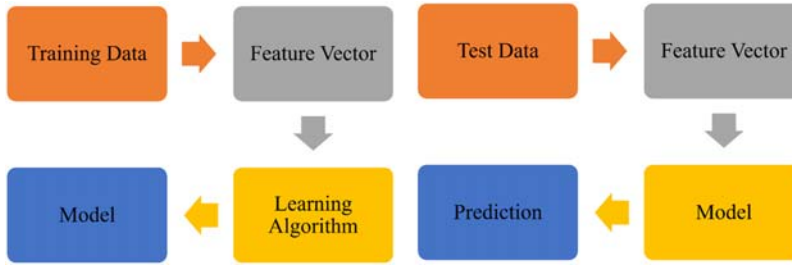

Figure 6.2 Data-driven modeling and training and model testing.

labels correspond to prediction using dataset and features as input [28]. Unsupervised learning is a branch of ML where we apply statistical learning methods to understand our data or create a better representation of it [28]. In this case, we do not have explicit labels.

In Fig. 6.2, a comprehensive ML-based modeling structure discussing the training and testing phase is shown. In continue, several methods, including classification and regression, association rules analytics, time-series analytics, behavioral and log analysis, etc., are explained and illustrated.

6.2.4.1 Classification

Classification is defined as the recognition and grouping process of objects into determined labels. By the precategorized training sets, classification in ML problems leverages many algorithms to classify future test sets into respective and appropriate categories. These classification methods utilize input training datasets aiming at the likelihood or probability prediction [29].

Some classification models are discussed in the following. First, Naive Bayes considers independent predictors, meaning that features are unrelated to each other. Second, the decision tree is used for visual decision-making representation. It is usually made by a yes/no question asking and splitting the answer to move to another decision. Third, k-nearest neighbor (KNN) is used for data division into classes based on the distance between the data points. This method assumes that close data points must be similar and the data points to be classified will be grouped with the closest cluster.

6.2.4.2 Regression

It is a supervised learning method to find the correlation between variables helping scientists with the continuous output variable prediction based on the independent variables. Regression's most popular use cases are forecasting, time-series modeling, and determination of the cause−effect relationship between variables. Simply, regression represents a curve (or a line) passing through all the data points on the target variable graph in such a way that the vertical distance between the data

points and the regression line is minimum. The distance between data points and curves interprets the strength of the model [29].

Some regression models are discussed in the following. First is a linear regression, which shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). The second is a polynomial regression, in which the original features are transformed into polynomial features of a given degree and then modeled using a linear model. The third is a support vector regression (SVR), which tries to determine a hyperplane with a maximum margin so that the maximum number of data points is covered in that margin. The SVR's main goal is to consider the maximum data points within the boundary lines and the hyperplane.

6.2.4.3 Clustering

It is a method in the field of unsupervised ML, which is known in many DS areas for statistical data analytics. Clustering algorithms are usually for the structure's recognition within a dataset. In the case of a lack of knowledge about groups and categories in the dataset, it will help with the classification of homogeneous groups of cases. In other words, identical data records are in a distinct cluster, which is different from other records in another cluster. Hence, various records sorting into homogeneous internally and heterogeneous externally groups are the most popular clustering goals. It even gives data scientists insight into how data is distributed in a given dataset or as a preprocessing phase for other algorithms.

Through the literature, K-means, hierarchical clustering, and CLARA are categorized into partitioning methods. Also, Density-Based Spatial Clustering of Applications with Noise and ordering points to identify the clustering structure (OPTICS) are categorized into density-based methods. The single and complete linkages are tools for implementing hierarchical methods. There are several categories left in this field: SG-based clustering algorithms, including STING, CLIQUE; model-based clustering such as neural networks, Gaussian mixture model, self-organizing map; and constrained-based algorithms including COP K-means, cyan, magenta, yellow, key (CMWK)-means, etc. [30].

6.2.4.4 Association rules learning

It is recognized as a rule-based learning system in the unsupervised learning area, which is mostly used for a relationship between features' establishment. Association rules learning is a descriptive method often used for large dataset analysis with the goal of pattern discovery. Its essential strength is its completeness and comprehensiveness, which is viable via user-operational constraints such as minimum support and confidence value.

It also helps a scientist to find and interpret trends and cooccurrences within large datasets. In an SG, for instance, the organization infers knowledge about the behavior of devices in terms of power, data flow in the system, etc. that will smooth the production plan change. Most known association rules methods are frequent

pattern-based, logic-based, tree-based, fuzzy rules, etc. In addition, the rule learning algorithms including AIS, Apriori, Apriori-TID, and Apriori-Hybrid, FP-Tree, Eclat, and RARM enable SGs for appropriate problem-solving. Apriori is the most popular implemented method for association rules discovery from a given dataset between all association rules learning methods [31].

6.2.4.5 Prediction and analytics for time-series data

When a dataset is indexed by the date and time stamp specifically including date and time stamp, it is called a time-series data type. From the frequency perspective, the time series are categorized into distinct types, including annual budget and expenses, monthly expenditure computation, and power and energy metrics (daily).

Mathematical modeling for time-series data and fitting process is known as time-series analysis. There are several popular time-series prediction algorithms being used for useful information extraction applications. For example, for time-series forecasting applications with attention to time patterns, the autoregressive (AR) model learns the behavioral trends or patterns of the historical data. Another algorithm is moving average (MA), which is a form of time-series smoothing and forecasting using historical errors in a regression model to interpret averaged trends [32].

The autoregressive moving average (ARMA) combines these two last methods, where AR extracts the momentum and pattern of the trend, and the MA captures the noise effects. The autoregressive integrated moving average (ARIMA) model is an extension of ARMA, which is a frequently used time-series model. As a generalization of an ARMA model, ARIMA is more adaptive than other statistical methods, that is, exponential smoothing and linear regression. The ARMA model is only available for stationary time series, even though the ARIMA model is usually being used for cases of nonstationarity as well. Likewise, seasonal autoregressive integrated moving average, autoregressive fractionally integrated moving average, and autoregressive moving average model with exogenous inputs model (ARMAX model) are other algorithms for time-series data [33]. Furthermore, ML-based and DL-based approaches are appropriate for effective time-series analytics, which is typically used for SG's power flow and consumption, manufacturing, event data, IoT devices data, and commonly in any applied science and engineering temporal measurement domain in SGs. All these time-series-based data analytics algorithms are shown in Fig. 6.3.

6.2.4.6 Behavioral data analysis

Behavioral data analysis helps with revealing new insights into SG sites and IoT applications, power flow behavior, and operators' behavior. It helps in pattern understanding and identifying the reason behind them enabling more accurate activities. Cohort analysis is a division of behavioral analytics that concentrates on studying behavior changes over time. Its popular methods are behavioral data clustering, behavioral decision tree classification, behavioral association rules, etc. [34].

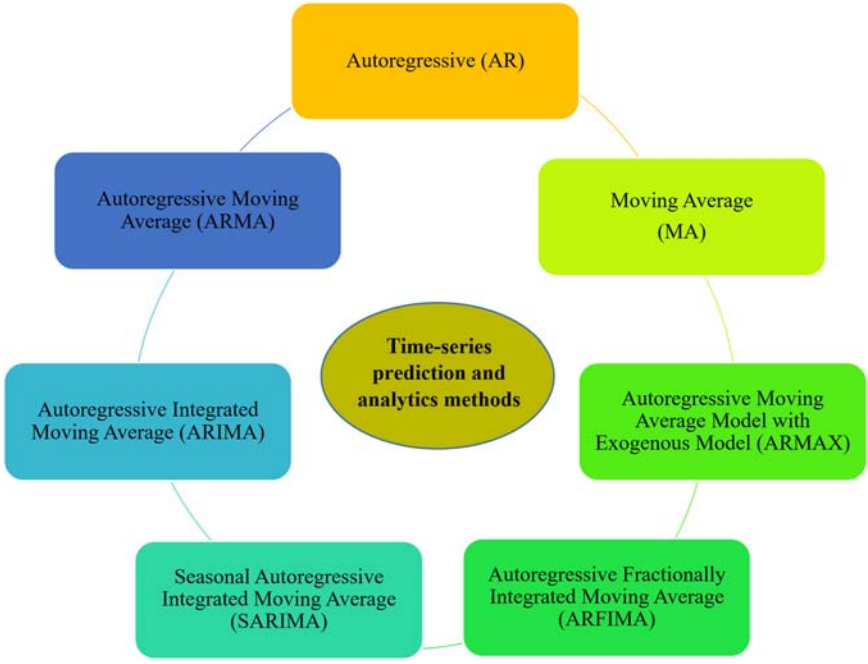

Figure 6.3 Time-series data prediction and analytics methods.

6.2.4.7 Anomaly detection

This section of AI is also known as outlier analysis, which is a DM phase detecting data points, and events that deviate from the regularities or normal behavior of a dataset. Anomalies are commonly recognized as outliers, noise, abnormalities, novelties, inconsistency, and exceptions. Exploring new situations and cases is viable via anomaly detection methods deviant based on historical data by interpreting data patterns using historical data.

Different data sources such as logs, facilities, the internet, and servers generate data that can be used in case of inconsistency removal in anomaly detection. While supervised learning, many ML algorithms including KNN, isolation forests, and clustering ease the anomaly exclusion process that results in a statistically substantial accuracy growth [35]. Nevertheless, extraction of useful variables, recognition of normal trends, imbalanced data management, addressing variations in abnormal behavior or irregularities, sparsity of abnormal events, circumstantial variations, etc. are challenging in the anomaly detection procedure. Anomaly detection applications are cyber-security analysis, intrusion, fraud, fault detection, and also ecosystem disturbance detection [36].

6.2.4.8 Factor analysis

Factors are considered as explanation of relationships or correlations between features. So, factor analysis (FA) is a learning method based on fundamental entities.

Its application is mainly feature organization by comparing them according to their total variance, with attention to mathematical and statistical procedures. The goal of FAis to determinethe features (factors) set number by degree calculation, so the fundamental impact underlying the data will be exposed. As a result, it will be clear which factors are associated with each other and also which factors contribute to the target variable. The ultimate purpose of FA is data summarization so that patterns between factors (variables) can easily be interpreted.

The two most common FA techniques are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Recognition of complex trends is possible by EFA analyzing the dataset and prediction evaluation. On the other hand, CFA's mission is to validate hypotheses and use path analysis figures for factor representation [37]. FA is categorized in unsupervised learning with the goal of dimensionality minimization. The most popular FA methods are principal components analysis, principal axis factoring, and maximum likelihood [38]. Also, for the quantification of the statistical interpretation between two continuous factors, some correlation analyses, that is, Pearson correlation analysis are being used.

6.2.4.9 Logs analytics

The data that record system runtime activities in detail and production patterns are known as logs. Log analytics is more of an interpretation approach, which is able to understand generated records/messages. Types of logs are system log, event log, device log, server log, network log, audit trail, record, etc. Data logging is defined as record creation. Many programmed technologies, such as network devices/servers and operating systems, generate logs. There are many smartphone devices in SGs that generate call logs, SMS/MMS Logs, app monitoring logs, notification logs, etc. In SGs, operators' actual behavioral activities with their devices are available as the main characteristics [39].

Appropriate log analytics tools are ML models, classification based on tagging, correlation analysis, and pattern recognition algorithms. Within SGs, log analysis helps with compliance with security procedures and power grids' regulations. It also finds the gaps in the system performance and so it supplies a better user experience by triggering the technical issue solving services. For example, log files are being used to record data on the internet between devices and among grid servers [40].

6.2.4.10 Deep learning and artificial neural networks

ANNs are used for building a computational architecture based on specific processing layers, including the input and output layers, and between them, the hidden layers. The most essential advantage of DL in comparison with regular ML methods is the performance from a complexity and time point of view, especially during the time of training on the large datasets. The most effective and popular DL methods are multilayer perceptron, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks [41]. There is also the "backpropagation"

technique, which adjusts the weights internally in the model structure building process. CNNs are developed on the basis of ANNs, consisting of convolutional layers, pooling layers with filters, and fully connected layers. Its use cases are natural language processing, speech enhancement and recognition, image processing, and autocorrelated data since the two-dimensional (2D) data type is the input of these cases. Advance CNN-based algorithms are AlexNet, Image-Net, Inceptions, Visual Geometry Groups (VGG), ResNet, etc. [42].

Furthermore, recurrent neural networks (RNNs) are another popular kind of DL network and some algorithms such as LSTM and GRU are categorized as RNN methods. Despite regular feed-forward neural networks, LSTM involves feedback through the network by the connections between units. Hence, LSTM networks are appropriate for sequential data analysis and interpretation, including tasks such as classification, sorting, and prediction for time-series data [43]. Thus, with the sequential data type, RNNs can be widely used.

6.3 Big data

A great amount of data created from different data sources in SGs are named "big data." Big data is mostly for managing and facilitating smart infrastructures in modern power systems. The big data issues in SGs are related to different sources such as including phasor measurement unit (PMU) data, power consumption records, advanced metering infrastructure measurements data, and power system smart meter monitoring, maintenance, and management data [44].

Information technology methods are now available in SGs for solving big data issues. Recent surveys have studied some big data challenges in SG systems including the big data role in technology and management in SGs and future issues [45]. Given the recent advancements in smart technologies, it is essential to study and review big data challenges in SGs with attention to SG-oriented viewpoint's theoretical, analytical, and standardized features. In this regard, the research gaps would be the need for the cooperative SG-oriented, information and communication technology−oriented approaches for both scientific communities' experts and public-sector SG administrators and operators as big data application users with limited information of big data challenges. In the following sections, the literature, application, and management approaches for big data are discussed.

6.3.1 Big data in smart grid literature

Based on various definitions of big data, the current big data can be compared to traditional data in order to clarify its applicableness. For the classification of big data characteristics, new methods are needed based on distinct definitions from typical data sources [46]. The dataset records quantity is rising daily by the scale of TBs and PBs because of the emerging big data. Classified structures recommended by big data instead of typical data structures in traditional data sources are

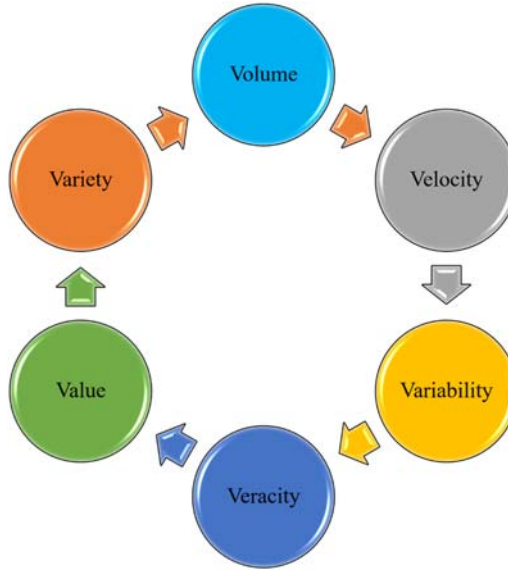

Figure 6.4 Big data 5Vs.

considered as another advantage. The big data velocity characteristic is mainly for computation of the processing rate in the dataset, which is really needed for the data streams and the real-time applications. At last, the goal of big data processing will be extraction of important values from the huge amount of data for real-time applications in SGs.

Volume (dataset size), velocity (the process of collecting big data and analyzing its speed and timeliness), variety (structured, semistructured, and unstructured heterogeneous data types), value (information extracted from big data sparsity), veracity (big data structures credibility and safety), and variability (data attribute modification) are the most essential attributes known as big data 5Vs [47] as shown in Fig. 6.4.

6.3.2 Big data architecture in smart grids

The big data system architecture can be demonstrated as a value chain consisting of four phases that are discussed in the following. While these four phases are defined and are under consideration, the raw data created by big data sources are analyzed and modified for gaining knowledge with attention to distinct management and control goals.

1. *Data generation*: Big data generation along with its types, characteristics, and origins from various data sources is the task of this phase.
2. *Data acquisition*: Aggregates of big data for effective outputs are discussed in this phase, including tasks such as data gathering (information retrieval from sensors and measurement facilities in SGs), preprocessing (errors integration, eliminating redundant records,

and compressing the results), filtering, features engineering, and transmission [transferring data into DCs for storage by communication infrastructures] [48].

3. *Data storage*: Big data storing and management for future applications are tasks for this phase.
4. *Data analysis*: Enhancement of the analytical aspects of the procedure to interpret the collected data by modeling methods for classification and other DS approaches discussed earlier. It is the most important phase in big data for making decisions based on big data knowledge in SGs. Big data analytics are also classified into descriptive, predictive, prognostic, and prescriptive analytics according to the depth of analysis, and their methods are classified into data visualization, and statistical analysis. [49].

6.3.3 Big data technologies in smart grids

The big data technology is utilized for a huge amount of data and the case that typical DS methods or other hybrid ML and DL algorithms are not capable to deal with big data architecture issues discussed earlier, that is, data acquisition and data analysis according to restrictions. Newly, big data technologies have emerged in modern grids, which play a critical role in developing an integrated system to supply efficient operations on data so that timely event detection will happen [50]. Big data tool integration is vital in SG applications. For example, state-of-the-art bidirectional communication and big data management technologies are substantial in SGs, that is, dispatchable energy resources containing electrical loads among themselves and development of distributed energy resources (DERs) [51]. PMU data is applicable in smart distribution grids regarding the demand response for power systems and DERs as practical resources of adaptability. The goal of big data technology could be an analytical framework that applies agent-based modeling for data processing purposes in the SG.

Popular big data technologies in SGs are DM, IoT, cloud computing, software-defined networking, and network function virtualization. DM is defined as pattern recognition by means of computational algorithms with approaches such as estimation, ML, and statistics, which are applicable in SGs [52]. Classification of the devices and other facilities performance variables is an example of DM use case in SGs, according to their power consumption type. Subsequently, AI methods are applicable for designing a risk model and solving uncertainty issues within load forecasting. These methods are developed with the fuzzy wavelet neural network to help in predicting power consumption and also a singular nonparametric estimation algorithm for the power distribution network scheduling [53].

Uncertainty and load-profile volatility, which are measured by smart metering data, are among the issues in SG load forecasting. Heterogeneous big data in the SG causes data-driven decision-making based on data streaming and load-profile behavior prediction with the help of ML and DL methods [54]. These methods are applied for grid energy management, safety, and security purposes. For instance, DL is able to enhance the AI performance through complicated data management, so SG administrators and data scientists are able to extract complex trends automatically, intelligent interaction with devices for maintenance and production, and more precise power load forecasting [55].

6.3.4 Big data tools in smart grids

Here, some of the useful tools for big data analysis, such as Apache Drill, Hadoop, game theory, and DCs, are discussed.

6.3.4.1 Apache Drill

The end users of SGs give the managers and administrators real-time interaction by computers and processing units helping the data to be generated in a more interactive environment. Apache Drill is mainly a distributed system providing interactive analysis for big data and its aim is to contain responses with low latency to ad hoc queries [56].

6.3.4.2 Hadoop

Hadoop is the most widespread big data technology applied in SGs, which provides network searching, click-stream analytics, and spam data filtering. The CloudView, which is a framework based on Hadoop, is for local load utilization, management of various architectures, and clusters of big data analysis purposes. Hadoop consists of two essential counterparts: first, Hadoop distributed file system for storing data according to the capacity prioritized by available tasks and second, the Hadoop execution engine for data processing purposes [57].

6.3.4.3 Game Theory

Game theory is a mathematical method for users' behavior analysis in many applications. In the literature, the multistage Stackelberg game is mentioned as an efficient tool for optimization problem-solving in the field of micro-SG energy management. Other AI complicated models including reinforcement learning, dynamic programming, and game theory are applicable in SG's security awareness management [58]. It works as a security-aware resource allocation model in a matching-coalition game scope.

6.3.4.4 Data centers

These tools represent platforms for dense storage of data enabling the system for multiple functionalities such as data acquisition, transmission, and processing. DCs are able to handle big data by maintaining requirements under time restrictions. The characteristics of DCs, including protection, extensibility, redundancy, energy effectiveness, and consistency, can maintain standardized functionality of big data features by an authorized infrastructure [59]. Cloud-based DCs are capable of power efficiency improvement and reduction in the cost of energy. According to the obvious rise in the data generation rate in power grids lately, it is expected that the big data challenges in DCs attract more scholars to pay attention to the potential topics in this regard, which may help to provide better insights for decision-making out of big data in DCs [60].

6.3.5 Big data applications in smart grids

The SG challenges are usually about data generation, preprocessing, distribution, end users, resources, devices, networks, and interaction with one another for applications and purposes. In this part, the big data systems' most important applications in SGs are discussed and summarized. These applications are expanded in distinct power system phases consisting of the power system scheduling, modeling, monitoring, control, security, and safety [61]. Power generation management is an essential application in SGs, which is adaptable for customizing the scheduling and facilitating the decision-making procedure. The great aspects of this application are efficient power load dispatch, the performance of power generation and storage systems, and power grid optimization in terms of production and cost [62].

Another application of big data is renewable energy resources (RERs) and microgrid management, which is a promising technology for forecasting and management improvement. Based on the literature, this technology is applicable to the association of RERs including wind, solar, biomass, and marine energies [63]. Microgrids known as the new integration of distributed power generation methods are considered as another big data application in SGs. In this regard, the most two important tasks are microgrid investment scheduling and microgrid load distribution optimization. Also, demand-side management is another big data technology in smart power systems.

6.4 Future research potentials

In this section, future possible fields of study in DS and big data analysis for IoT energy systems are discussed. First, future challenges should be presented. Herein, SG development needs to bring up and categorize challenges.

6.4.1 Security and privacy

The security issue of SGs is one of the main challenges in the future. In the literature, attacks and defenses on DL models have been detected, which influence models' performance and make the accuracy, credibility, and trustworthiness less. One of a kind of these threats is false data entrance, which enables IoT devices and sensors with different configurations to send false data. This threat causes faulty results, recommendations, and prognosis to the analytics operations.

Also, SGs are dealing with privacy challenges in the context of IoT applications. In this regard, the data that are captured every second in different locations of the grid by the sensors may contain distinct layers of information. This information affects the decision-making process for different senior managers and middle managers. First, sending and receiving sensitive data to servers with different authentications for administrators presents several privacy concerns. The second challenge in the field of privacy will be the loss of data. Third, the SG's database may be exploited by a third party who does not have permission to employ or even access

the information. However, even the current solutions will remain vulnerable and are able to be hacked by robust attacks. In fact, the hacking process is the development of DL algorithm that learns other DL methods' threat detection methods. This process is explained by how these DL methods generate attacks that are difficult to detect.

6.4.2 Internet of Things big data challenges

Due to the great volume of data generated by different IoT devices, several issues are being concerned, including data depository, communication, complexity, and data analysis. The generated data storage for a long time introduces an important challenge because of the lack of ability to be managed using traditional database management tools. Hence, modern specific tools and infrastructures are essential for handling structured and unstructured big data. Furthermore, the IoT big data analytics requires specific technologies for extracting valuable insights including efficient high-tech processors, levelized edge computing tools, and modern, high-quality, and quick software for big data analysis as discussed before, that is, Apache Hadoop.

6.4.3 Deep learning implementation challenges and limitations

Although satisfying results have been achieved by DL models in SGs' data analytics, in some situations these models are not adequate solutions. Several restrictions should be considered; First, a large amount of data is required for providing good results by DL models. Second, the more data is gathered for the database, the more the training process complexity will appear, which is computationally expensive, and extremely time-consuming. Third, limited access to datasets suffers SG transformation scientists since they have to develop specific datasets requiring a great deal of time and effort. Fourth, DL models are not appropriate for similar but not identical purposes. Fifth, in a black-box running process, the model results' accuracy improvement occurs with time, or it may prove a vulnerable point, especially in the prediction process. In black-box models like neural networks, assuring the training performance is difficult and in some situations, the models' results are not effective.

6.4.4 Smart grid data–driven planning, cost management, and quality of service

A SG IoT system construction needs scheduling strategies that must be considered before launching. A development plan must be designed for each section of the grid based on the needs of its devices because avoiding problems and faults including redundant services or uncoordinated power networks will be solved by this planning. Appropriate planning helps with the facilities development and grid services by identifying the required areas and facilitating the legacy systems integration within the new system.

In this regard, the main SG characteristic is the interconnection of the components, which is considered as an expensive service for governments where many devices, sensors, actuators, and software should be deployed. For instance, a smart surveillance system installation needs modern systems and platforms to gather and process data. These prerequisites are expensive to implement and any error often causes high losses. For maintaining a successful SG system, the quality of service (QoS) is an essential characteristic of applications that should be certified. Separate QoS measurements are used for quality examination such as response time, availability, extensibility, and trustworthiness. Many technologies are integrated for SG service development, such as hosting services and storage frameworks. The assurance of QoS provided by the discussed technologies to ensure an adaptable, powerful, and trustworthy energy system is vital.

6.5 Conclusion

Knowledge-based tools such as DS and big data analysis aid us to comprehend the data's nature and use the extracted insight for improving systems' optimization. How to automize the process for working in a more optimized way according to analytics and algorithms used for performance evaluation is important to consider. The data gathering process from energy systems needs to be improved by experienced experts to elevate the performance results. All features that are available by the data gathering should be revised by the DS specialists, so some first empirical studies can help us interpret the data at first and then find the structure to do the related analytics. This analysis will be much different in case of having a large number of transactions (rows) meaning that we are dealing with big data. Some specific distributions and platforms will be used to handle this data and tasks will be scheduled with network and computing-based methods using nodes and connections.

The integration of information and communication in the energy supply chain deploying IoT occurs with a variety range of embedded actuators, devices, and sensors. Although the usage potential of the internet is now appropriate for reaching the energy system devices attaining by standardized communication protocols, IoT in energy system implementation has several challenges, including accessibility, security, controlling the bandwidth, appropriate interface, connectivity, packet loss, and data analyzing. Available and insightful information is being used for the decision-making process in energy systems, which ends in big data according to great data log, namely, vibration, temperature, and flow sensors outputs. All the gathered information will be analyzed in the big data format. The result will be a more accurate prediction, optimized performance, and better fault diagnosis in energy devices.

References

[1] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, The rise of "big data" on cloud computing: review and open research issues, Inf. Syst

47 (2015) 98−115. Available from: https://doi.org/10.1016/j.is.2014.07.006. ISSN 0306-4379.

[2] S. Pradeep, J.S. Kallimani, "A survey on various challenges and aspects in handling big data," in: 2017 International Conference on Electrical, Electron. Communication, Computer, Optim. Tech. (ICEECCOT), 2017, pp. 1−5, https://doi.org/10.1109/ICEECCOT.2017.8284606.

[3] L. Dobrica, E. Niemela, "A survey on software architecture analysis methods," in, IEEE Trans. Softw. Eng 28 (7) (2002) 638−653. Available from: https://doi.org/10.1109/TSE.2002.1019479.

[4] O. Rusu, et al., "Converting unstructured and semi-structured data into knowledge," in: 2013 11th RoEduNet International Conference, 2013, pp. 1−4, https://doi.org/10.1109/RoEduNet.2013.6511736.

[5] K. Zhou, C. Fu, S. Yang, Big data driven smart energy management: from big data to big insights, Renew. Sustain. Energy Rev 56 (2016) 215−225. Available from: https://doi.org/10.1016/j.rser.2015.11.050. ISSN 1364-0321.

[6] I.K. Nti, J.A. Quarcoo, J. Aning, G.K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, prospects," Big Data Min. Analytics 5 (2) (2022) 81−97. Available from: https://doi.org/10.26599/BDMA.2021.9020028.

[7] A. Belhadi, K. Zkik, A. Cherrafi, S.M. Yusof, S. El Fezazi, Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies, Computers & Ind. Eng 137 (2019) 106099. Available from: https://doi.org/10.1016/j.cie.2019.106099. ISSN 0360-8352.

[8] C. Sarra, Data mining and knowledge discovery. Preliminaries for a critical, examination of the data driven society, Glob. Jurist 20 (1) (2020) 20190016. Available from: https://doi.org/10.1515/gj-2019-0016.

[9] A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manag 35 (2) (2015) 137−144. Available from: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. ISSN 0268-4012.

[10] C.L. Stergiou, A.P. Plageras, K.E. Psannis, B.B. Gupta, in: B. Gupta, G. Perez, D. Agrawal, D. Gupta (Eds.), Secure Machine Learning Scenario from Big Data in Cloud Computing via Internet of Things Network, Handbook of Computer Networks and Cyber Security. Springer, Cham, 2020. Available from: https://doi.org/10.1007/978-3-030-22277-2_21.

[11] O. Müller, I. Junglas, J. Brocke, et al., Utilizing big data analytics for information systems research: challenges, promises and guidelines, Eur. J. Inf. Syst 25 (2016) 289−302. Available from: https://doi.org/10.1057/ejis.2016.2.

[12] D.M. Dimiduk, E.A. Holm, S.R. Niezgoda, Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, Processes, Struct. Engineering. Integr. Mater. Manuf. Innov 7 (2018) 157−172. Available from: https://doi.org/10.1007/s40192-018-0117-8.

[13] S. Ewa, Modern data science for analytical chemical data − a comprehensive review, Analytica Chim. Acta 1028 (2018) 1−10. Available from: https://doi.org/10.1016/j.aca.2018.05.038. ISSN 0003-2670.

[14] C. Longbing, Data science: a comprehensive overview, ACM Comput. Surv. 50 (3) (2017) 42, Article 43. Available from: https://doi.org/10.1145/3076253.

[15] A. Kumar, S.R. Sangwan, A. Nayyar, Multimedia social big data: mining, in: S. Tanwar, S. Tyagi, N. Kumar (Eds.), Multimedia Big Data Computing for IoT Applications. Intelligent Systems Reference Library, 163, Springer, Singapore, 2020. Available from: https://doi.org/10.1007/978-981-13-8759-3_11.

[16] B.T. Hazen, J.B. Skipper, C.A. Boone, et al., Back in business: operations research in support of big data analytics for operations and supply chain management, Ann. Oper. Res 270 (2018) 201−211. Available from: https://doi.org/10.1007/s10479-016-2226-0.
[17] D. Alahakoon, X. Yu, Smart electricity meter data intelligence for future energy systems: a survey, IEEE Trans. Ind. Inform. 12 (1) (2016) 425−436. Available from: 10.1109/TII.2015.2414355.
[18] Y. Claudia Vitolo, D. Elkhatib, C.J.A. Reusser, W.Buytaert Macleod, Web technologies for environmental Big Data, Environ. Model. & Softw 63 (2015) 185−198. Available from: https://doi.org/10.1016/j.envsoft.2014.10.007. ISSN 1364-8152.
[19] C.-H. Cheng, Y.-F. Kao, H.-P. Lin, A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes, Appl. Soft Comput 108 (2021) 107487. Available from: https://doi.org/10.1016/j.asoc.2021.107487. ISSN 1568-4946.
[20] F. Steven, Multiple classifier architectures and their application to credit risk assessment, Eur. J. Operational Res 210 (2) (2011) 368−378. Available from: https://doi.org/10.1016/j.ejor.2010.09.029. ISSN 0377-2217.
[21] A.A. Solanke, M.A. Biasiotti, Digital forensics AI: evaluating, standardizing and optimizing digital evidence mining techniques, Künstl. Intell. (2022). Available from: https://doi.org/10.1007/s13218-022-00763-9.
[22] S. Ren, Y. Zhang, T. Sakao, et al., An advanced operation mode with product-service system using lifecycle big data and deep learning, Int. J. Precis. Eng. Manuf.-Green Tech 9 (2022) 287−303. Available from: https://doi.org/10.1007/s40684-021-00354-3.
[23] M. Veale, R. Binns, Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data, Big Data & Soc (2017). Available from: https://doi.org/10.1177/2053951717743530.
[24] M. Marathe, K. Toyama, Semi-automated coding for qualitative research: a user-centered inquiry and initial prototypes. in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, 2018 Paper 348, pp. 1−12. https://doi.org/10.1145/3173574.3173922.
[25] S.G. Heeringa, B.T. West, P.A. Berglund, Applied Survey Data Analysis, 2nd ed., Chapman and Hall/CRC, 2017. Available from: https://doi.org/10.1201/9781315153278.
[26] M. Seyedan, F. Mafakheri, Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities, J. Big Data 7 (2020) 53. Available from: https://doi.org/10.1186/s40537-020-00329-2.
[27] E.T. Bradlow, M. Gangwar, P. Kopalle, S. Voleti, The role of big data and predictive analytics in retailing, J. Retail 93 (1) (2017) 79−95. Available from: https://doi.org/10.1016/j.jretai.2016.12.004. ISSN 0022-4359.
[28] M. Clayton, N. Zoltán, S. Arno, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev 81 (1) (2018) 1365−1377. Available from: https://doi.org/10.1016/j.rser.2017.05.124. ISSN 1364-0321.
[29] C. Kuster, Y. Rezgui, M. Mourshed, Electrical load forecasting models: a critical systematic review, Sustain. Cities Soc 35 (2017) 257−270. Available from: https://doi.org/10.1016/j.scs.2017.08.009. ISSN 2210-6707.
[30] M.A. Mahdi, K.M. Hosny, I. Elhenawy, Scalable clustering algorithms for big data: a review, IEEE Access 9 (2021) 80015−80027. Available from: https://doi.org/10.1109/ACCESS.2021.3084057.

[31] M. Hahsler, R. Karpienko, Visualizing association rules in hierarchical groups, J. Bus. Econ. 87 (2017) 317−335. Available from: https://doi.org/10.1007/s11573-016-0822-8.
[32] V. Prema, K. Uma Rao, Development of statistical time series models for solar power prediction, Renew. Energy 83 (2015) 100−109. Available from: https://doi.org/10.1016/j.renene.2015.03.038. ISSN 0960-1481.
[33] F. Saâdaoui, H. Rabbouch, A wavelet-based hybrid neural network for short-term electricity prices forecasting, Artif. Intell. Rev 52 (2019) 649−669. Available from: https://doi.org/10.1007/s10462-019-09702-x.
[34] I.H. Sarker, Context-aware rule learning from smartphone data: survey, challenges and future directions, J. Big Data 6 (2019) 95. Available from: https://doi.org/10.1186/s40537-019-0258-4.
[35] G. Padmavathi, D. Shanmugapriya, A. Roshni, "Performance analysis of unsupervised machine learning methods for mobile malware detection," in: 2022 9th International Conference on Computing for Sustainable Global, Dev. (INDIACom), 2022, pp. 201−206, https://doi.org/10.23919/INDIACom54597.2022.9763180.
[36] M. Ahmed, A. Naser, J. Mahmood, Hu, A survey of network anomaly detection techniques, J. Netw. Computer Appl 60 (2016) 19−31. Available from: https://doi.org/10.1016/j.jnca.2015.11.016. ISSN 1084-8045.
[37] N. David Bowman, A.K. Goodboy, Evolving considerations and empirical approaches to construct validity in communication science, Ann. Int. Commun. Assoc 44 (3) (2020) 219−234. Available from: https://doi.org/10.1080/23808985.2020.1792791.
[38] R. Bhatia, S. Benno, J. Esteban, T.V. Lakshman, J. Grogan, Unsupervised machine learning for network-centric anomaly detection in IoT. in: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks (Big-DAMA '19). Association for Computing Machinery, New York, NY, USA, 2019, pp. 42−48. https://doi.org/10.1145/3359992.3366641.
[39] J. Li, H. Li, W. Umer, H. Wang, X. Xing, S. Zhao, et al., Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology, Autom. Constr 109 (2020) 103000. Available from: https://doi.org/10.1016/j.autcon.2019.103000. ISSN 0926-5805.
[40] K. Guo, Y. Lu, H. Gao, R. Cao, Artificial intelligence-based semantic internet of things in a user-centric smart city, Sensors 18 (2018) 1341. Available from: https://doi.org/10.3390/s18051341.
[41] N. Lingling, D. Wang, V.P. Singh, J. Wu, Y. Wang, Y. Tao, J. Zhang, Streamflow and rainfall forecasting by two long short-term memory-based models, J. Hydrol 583 (2020) 124296. Available from: https://doi.org/10.1016/j.jhydrol.2019.124296. ISSN 0022-1694.
[42] N. Wang, Y. Wang, M.J. Er, Review on deep learning techniques for marine object recognition: architectures and algorithms, Control. Eng. Pract 118 (2022) 104458. Available from: https://doi.org/10.1016/j.conengprac.2020.104458. ISSN 0967-0661.
[43] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time series through stochastic recurrent neural network. in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, 2019, pp. 2828−2837. https://doi.org/10.1145/3292500.3330672.
[44] H. Akhavan-Hejazi, H. Mohsenian-Rad, Power systems big data analytics: an assessment of paradigm shift barriers and prospects, Energy Rep 4 (2018) 91−100. Available from: https://doi.org/10.1016/j.egyr.2017.11.002. ISSN 2352-4847.

[45] E. Hossain, I. Khan, F. Un-Noor, S.S. Sikander, M.S.H. Sunny, Application of big data and machine learning in smart grid, and associated security concerns: a review, IEEE Access 7 (2019) 13960−13988. Available from: https://doi.org/10.1109/ACCESS.2019.2894819.

[46] L. Kong, Z. Liu, W. Jianguo, A systematic review of big data-based urban sustainability research: State-of-the-science and future directions, J. Clean. Prod 273 (2020) 123142. Available from: https://doi.org/10.1016/j.jclepro.2020.123142. ISSN 0959−6526.

[47] K. Adnan, R. Akbar, An analytical study of information extraction from unstructured and multidimensional big data, J. Big Data 6 (2019) 91. Available from: https://doi.org/10.1186/s40537-019-0254-8.

[48] T.A. Alghamdi, N. Javaid, A survey of preprocessing methods used for analysis of big data originated from smart grids, IEEE Access 10 (2022) 29149−29171. Available from: https://doi.org/10.1109/ACCESS.2022.3157941.

[49] H.-N. Dai, H. Wang, G. Xu, J. Wan, M. Imran, Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies, Enterp. Inf. Syst 14 (9-10) (2020) 1279−1303. Available from: https://doi.org/10.1080/17517575.2019.1633689.

[50] X. Yu, Y. Xue, Smart grids: a cyber−physical systems perspective, Proc. IEEE 104 (5) (2016) 1058−1070. Available from: https://doi.org/10.1109/JPROC.2015.2503119.

[51] P.D. Diamantoulakis, V.M. Kapinas, G.K. Karagiannidis, Big data analytics for dynamic energy management in smart grids, Big Data Res 2 (3) (2015) 94−101. Available from: https://doi.org/10.1016/j.bdr.2015.03.003. ISSN 2214-5796.

[52] L. Briceno-Mena, M. Nnadili, M. Benton, J. Romagnoli, Data mining and knowledge discovery in chemical processes: effect of alternative processing techniques, Data-Centric Eng 3 (2022) E18. Available from: https://doi.org/10.1017/dce.2022.21.

[53] M. Rafiei, T. Niknam, J. Aghaei, M. Shafie-Khah, J.P.S. Catalão, Probabilistic load forecasting using an improved wavelet neural network trained by generalized extreme learning machine, IEEE Trans. Smart Grid 9 (6) (2018) 6961−6971. Available from: https://doi.org/10.1109/TSG.2018.2807845.

[54] A. Kumari, S. Tanwar, Secure data analytics for smart grid systems in a sustainable smart city: Challenges, solutions, and future directions, Sustain. Computing: Inform. Syst 28 (2020) 100427. Available from: https://doi.org/10.1016/j.suscom.2020.100427. ISSN 2210-5379.

[55] J.M. Górriz, J. Ramírez, A. Ortíz, F.J. Martínez-Murcia, F. Segovia, J. Suckling, et al., Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications, Neurocomputing 410 (2020) 237−270. Available from: https://doi.org/10.1016/j.neucom.2020.05.078. ISSN 0925-2312.

[56] L. Ordonez-Ante, T. Vanhove, G. Van Seghbroeck, T. Wauters, F. De Turck, "Interactive querying and data visualization for abuse detection in social network sites," in: 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST), 2016, pp. 104−109, https://doi.org/10.1109/ICITST.2016.7856676.

[57] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, F. Herrera, Big data: tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce, Inf. Fusion 42 (2018) 51−61. Available from: https://doi.org/10.1016/j.inffus.2017.10.001. ISSN 1566-2535.

[58] Z. Zhang, D. Zhang, R.C. Qiu, Deep reinforcement learning for power system applications: An overview, CSEE, J. Power Energy Syst. 6 (1) (2020) 213−225. Available from: https://doi.org/10.17775/CSEEJPES.2019.00920.

[59] X. Chu, S. Nazir, K. Wang, Z. Leng, W. Khalil, Big data and its V's with IoT to develop, sustainability, Sci. Program (2021). Available from: https://doi.org/10.1155/2021/3780594. Article ID 3780594, 16 pages, 2021.

[60] S. Talwar, P. Kaur, S.F. Wamba, A. Dhir, Big data in operations and supply chain management: a systematic literature review and future research agenda, Int. J. Prod. Res 59 (11) (2021) 3509−3534. Available from: https://doi.org/10.1080/00207543.2020.1868599.

[61] S.M. Nosratabadi, R.-A. Hooshmand, E. Gholipour, A comprehensive review on microgrid and virtual power plant concepts employed for distributed energy resources scheduling in power systems, Renew. Sustain. Energy Rev 67 (2017) 341−363. Available from: https://doi.org/10.1016/j.rser.2016.09.025. ISSN 1364-0321.

[62] A. Chauhan, R.P. Saini, A review on integrated renewable energy system based power generation for stand-alone applications: configurations, storage options, sizing methodologies and control, Renew. Sustain. Energy Rev 38 (2014) 99−120. Available from: https://doi.org/10.1016/j.rser.2014.05.079. ISSN 1364-0321.

[63] A. Yadav, N. Pal, J. Patra, et al., Strategic planning and challenges to the deployment of renewable energy technologies in the world scenario: its impact on global sustainable development, Env. Dev. Sustain 22 (2020) 297−315. Available from: https://doi.org/10.1007/s10668-018-0202-3.

This page intentionally left blank