Machine Learning with Applications 9 (2022) 100363

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Renewable energy management in smart grids by using big data analytics and machine learning

Noha Mostafa [a,b,*], Haitham Saad Mohamed Ramadan [c,d], Omar Elfarouk [a]

[a] Mechanical Engineering Department, Faculty of Engineering, The British University in Egypt, 11837, Shorouk, Egypt
[b] Industrial Engineering Department, Faculty of Engineering, Zagazig University, 44519, Zagazig, Egypt
[c] Electrical Power and Machines Department, Faculty of Engineering, Zagazig University, 44519, Zagazig, Egypt
[d] ISTHY, Institut International sur le Stockage de l'Hydrogene, 90400, Meroux-Moval, France

ARTICLE INFO

Keywords:
Energy internet
Renewable energy
Smart grid
Big data analytics
Machine learning
Predictive models

ABSTRACT

The application of big data in the energy sector is considered as one of the main elements of Energy Internet. Crucial and promising challenges exist especially with the integration of renewable energy sources and smart grids. The ability to collect data and to properly use it for better decision-making is a key feature; in this work, the benefits and challenges of implementing big data analytics for renewable energy power stations are addressed. A framework was developed for the potential implementation of big data analytics for smart grids and renewable energy power utilities. A five-step approach is proposed for predicting the smart grid stability by using five different machine learning methods. Data from a decentralized smart grid data system consisting of 60,000 instances and 12 attributes was used to predict the stability of the system through three different machine learning methods. The results of fitting the penalized linear regression model show an accuracy of 96% for the model implemented using 70% of the data as a training set. Using the random forest tree model has shown 84% accuracy, and the decision tree model has shown 78% accuracy. Both the convolutional neural network model and the gradient boosted decision tree model yielded 87% for the classification model.

The main limitation of this work is that the amount of data available in the dataset is considered relatively small for big data analytics; however the cloud computing and real-time event analysis provided was suitable for big data analytics framework. Future research should include bigger datasets with variety of renewable energy sources and demand across more countries.

1. Introduction

The increasing need for energy made it inevitable to resort to renewable sources. For years, many power companies have been installing renewably energy power stations worldwide to provide economic and clean energy (Missaoui et al., 2014). Renewable energy such as wind turbines and solar power have many advantages like low delivery costs and less emissions. However, traditional designs of grid energy storage systems are becoming impractical. Large blackouts that occur from time to time have highlighted the necessity to have an improved decision-making process that requires timely and accurate data on the dynamic events, the operating conditions and the sudden changes in the power. According to Zhou and Yang (2016), since the second industrial revolution, energy systems have passed four stages; decentralized systems, centralized systems, distributed systems and the most recent one, smart and connected systems or 'Energy Internet' that depends on innovative technologies such as mobile applications, Internet of Things (IoT), big data analytics (BDA), and cloud computing. Rifkin (2011) has defined energy internet as new energy utilization system that integrates renewable energy sources, distributed power stations, hydrogen energy, storage technologies and electric vehicles with the Internet technologies. The author has defined four characteristics for energy Internet; Powered by renewable energy sources, supports access to large-scale generation and storage systems, supports energy sharing, and supports the electrification of transportation systems.

Unlike fuel-based energy power stations, renewable energy requires more advanced management of power, balancing, and production capacity, which can be achieved by using smart grids (Rathor & Saxena, 2020). These grids integrate traditional power grids with advanced Information Technology (IT) and communication networks to deliver electricity with improved efficiency and reliability, while reducing cost and environmental impacts (Yan et al., 2013). Renewable energy resources are one of the major smart grid enablers in the residential neighborhoods, transformers and substations (Tene & Polonetsky, 2013). They can supplement power sources that can be quickly

* Corresponding author at: Mechanical Engineering Department, Faculty of Engineering, The British University in Egypt, 11837, Shorouk, Egypt.
E-mail addresses: noha.mostafa@bue.edu.eg, namostafa@eng.zu.edu.eg (N. Mostafa), haitham.mohamed-ramadan@utbm.fr (H.S.M. Ramadan), Omar.elfarouk@bue.edu.eg (O. Elfarouk).

https://doi.org/10.1016/j.mlwa.2022.100363
Received 14 February 2022; Received in revised form 13 June 2022; Accepted 13 June 2022
Available online 17 June 2022

installed, monitored and controlled for being used during the peak hours. Such resources must be observed carefully to capture all the possible opportunities for harvesting energy and for responding to any abnormalities. In the era of IT, many tools and machines are used for monitoring and controlling these sources. Examples on these IT tools include cyber–physical systems (Lee et al., 2014), distribution management system, geographic information systems, outage management systems, customer information systems, and Supervisory Control and Data Acquisition system (SCADA) (Al-Ali & Aburukba, 2015).

Managing these grids requires efficient real-time data processing and analysis of the massive amount of data captured by monitors, sensors, meters, cameras and computers to both improve the efficiency and to avoid delays and system shutdowns (Chen et al., 2014). This data is used for many different applications: real-time vulnerability assessment, demand management, predictive analytics, theft detection, energy trading, economic dispatch, etc. (Asad & Chaudhry, 2017). It is considerably difficult to manage these large and diverse datasets through using traditional database management tools; this type of data, namely 'Big data' requires advanced management approaches. Big data is a computer science technology that can be applied to data from different and uncorrelated sources, which make it hard to use traditional data analysis tools to process this data. It is characterized by its four main components: Variety, Velocity, Volume and Veracity, namely 'the 4Vs' (Wang et al., 2019). The 'Varity' component refers the different sources and types of data to be processed. The 'Velocity' component refers to the need for fast and synchronized processing and analysis of data. The 'Volume' component refers to the ability to handle large and growing amount of data (Sagiroglu & Sinanc, 2013). Finally, the 'Veracity' component is concerned with the uncertainty of data processing and the poor data quality (Kepner et al., 2014).

Over the past few years, big data tools have been adopted by many major companies in the power industry such as IBM, Siemens, General Electric and Oracle (Arenas-Martinez et al., 2010). According to Li et al. (2012), about 85% of the processing tasks of big data can be delayed by a day. Hence, even if the energy output is time-varying and intermittent, it can be leveraged for processing the delayed datasets. Every day, more Terabytes of data emerge at the energy data center. Hence, it has become crucial to adopt big data technology to extract information from the multiple and diverse data sources through novel data centers (Liu et al., 2012). Through BDA, it would be possible to understand the behavior of energy consumption, which would help to improve energy efficiency and also promote the concepts of sustainability (Koseleva & Ropaite, 2017; Marmaras et al., 2017; Mostafa & Negm, 2018). However, there are many challenges related to the market sector such as: the availability of building data on a large scale, the lack of data format and data field definitions for building datasets and the noise and uncertainty of the available empirical data (Mathew et al., 2015). Another important application is the use of BDA for energy management in smart cities (Strohbach et al., 2015). Other recent works have employed modeling and optimization techniques for renewable energy (Ju et al., 2021, 2019). From these works it was suggested that the self-adjustment mechanism can boost the performance of the traditional genetic algorithm, as it can conduct a self-check and find the worst solutions and relocate to better locations sampled from the Support vector regression (SVR) surfaces.

The research problem under consideration is to provide the maximum stable smart grid system through identifying the stability of a complex smart grid system consisting of many renewable energy input sources and many outputs.

The objective of this study is to explore the potential of using BDA in smart grids based renewable energy power stations, and to review the previous studies regarding this issue. Also, data analytics algorithms are applied on a big dataset and results were analyzed in terms of stability and accuracy. The rest of this paper is organized as follows: In Section 2, a review on the recent literature on using big data technology in renewable energy networks and smart grids is given. Section 3 discusses the technologies and theories used in BDA. In Section 4, a framework is proposed to link the architecture of big data with the smart grids and show the potential applications of the yielded information. In Section 5, an approach is proposed for predicting the smart grid stability based on BDA and machine learning. Results are summarized and discussed in Section 6, and finally, conclusions and future recommendations are given in Section 7.

2. Literature review

Over the past few years, many architectures and frameworks have been suggested to handle the data problems in power stations and smart grids.

2.1. Analytical models for managing power systems

In Billinton and Gao (2008), analytical models were used to assess wind energy generating system, Monte Carlo state sampling techniques have shown that five-state model of wind energy conversion system can provide an efficient assessment of the power system adequacy studies. The problem of temporally variability in renewable electric production was addressed in Goyena et al. (2009). For this purpose, the energetic balance algorithm has been considered to warranty the electric demand in large energy storage system. This algorithm has been used for analyzing the measured data constantly and accurately, and based on the results, data storage level can be determined. In Rogers et al. (2010), a framework for smart grids was proposed to restore system voltage while coordinating multiple power devices by following a chain of command structures to respond quickly and effectively to low-voltage incidents. The challenge was in the complexity of adding more devices and functions, the proposed system can provide wide, secure and versatile control of the smart grid. In Aquino-Lugo et al. (2011), agent-based technologies were used to manage data processing in smart grids. In Khalid et al. (2019), fuzzy logic and heuristics were used for energy management and control of home appliances with three criteria: cost, user comfort, and peak-to-average ratio.

In Labeeuw and Deconinck (2013), Markov models were used to build a behavior model based on the probability distribution of the electrical power. Then, the customers have been grouped into clusters according to their energy consumption and transferred into load profile documents. In Rahimi-Eichi et al. (2015), the effect of accurate prediction of driving speed on range estimation was evaluated. The driving speed profile has been regenerated using Markov chain model based on historical data. Luo and Oyedele (2022) have developed an integrated model based on self-adaptive deep learning model and particle swarm optimization to predict residential electricity load with moving horizons. The proposed model has shown its accuracy and robustness with a coefficient of determination up to 98.9%.

2.2. Using big data analytics in energy management applications

One of the earliest works in using BDA in energy management was in Su and Chow (2012), where big data theories were used to develop an estimation of distribution algorithm to allocate electric energy to electric vehicles at municipal parking lot. However, the increase of the number of vehicles connected to the grid has led to a challenging study with effect on quality and stability of the overall system. In Kwac and Rajagopal (2013), a methodology was developed for large-scale consumer targeting by combining BDA and scalable selection procedures via stochastic knapsack problem and demand response modeling. Accordingly, the fast heuristic algorithm has been considered to cope with computational issues resulting from the big volume of dataset. Different techniques have been used in big data processing including optimization, statistical models and data mining to handle big data of large volume within limited time. In Rahimi-Eichi and Chow (2014), a BDA framework was proposed to estimate the driving range of

Table 1
Energy consumption levels in data centers over different years (Avgerinou et al., 2017).

Region	Year	Consumption (TWh)
EU	2000	18.3
	2005	41.3
	2007	56
	2010	72.5
	2020	104
US	2013	91
	2020	140
Global	2007	216
	2012	269

electric vehicles. The framework has collected historical and real-time data from different sources, and then analyzes them through the range estimation algorithm.

A high-level look was given in Tannahill and Jamshidi (2014) at some Matlab tools that enable information extraction from big data sources. The proposed model has enabled predicting the amount of solar power generated by the micro-grid. Several techniques were experimented to reduce data dimensions and maximize the retained information with minimal error; these tools included dataset sanitation, input parameter selection, model generation via fuzzy clustering and rule inference, and neural network with back propagation. In Diamantoulakis et al. (2015), the use of big data techniques for dynamic energy management in smart grid platforms was addressed focusing on smart grid data mining, predictive analytical methods and smart meter data. The authors have argued that the most important challenge is the users' participation in cost reduction. In He et al. (2017), architecture was proposed as a data driven solution to detect abnormalities. Random matrix theory has been used to model this architecture and to conduct high-dimensional analysis and distributed calculation of the system.

In Kung and Wang (2015), an integrated system was proposed to combine renewable energy resources with cost benefit and big data analysis. The continuous Markov chain modeling has been used for analyzing historical electricity data in random time by employing time series analysis and multi-objective models. The objective of this system was to support strategic decisions on renewable energy investment and combination through these models, and to construct an enterprise-oriented cloud system and user interface. In Mashayekhy et al. (2014), MapReduce, a programming model for processing and generating big datasets, and its open-source implementation Hadoop were used. Two fast and energy aware scheduling algorithms have been used for real-time situations and performed simulation to analyze the performance of the proposed algorithms. The results showed the adequate capability of the algorithms to obtain near optimal solutions which leads to significant energy savings. In Pan et al. (2016), a smart metering methodology was proposed that can be used for BDA applications based on characteristic consumer load shapes. The data from smart meters can enable several applications such as forecasting, pricing, load profiling, connecting renewable energy to grids and irregularities capturing. A recent work by Zhang et al. (2021) has proposed a model based on convolutional neural network and Sequence-to-Sequence to perform multi-task learning for short-time multi-energy load forecasting. The results have shown that the model can effectively extract the overall feature and the time series feature.

2.3. Sustainability considerations in energy management

Large scale computational systems consume big amount of energy as data volume increases and accordingly more analysis is required. Table 1 shows the energy consumption levels in data centers in EU, USA and Globally over different years expressed in terawatt-hour (TWh) (Avgerinou et al., 2017).

From Table 1, it can be seen that there is a big increase in energy consumption in data centers. Hence, more advanced mechanisms are required for power control and management without decreasing accuracy and accountability. Green data centers, powered by renewable energy, have recently received an increasing academic interest from economic and environmental perspectives (Berral et al., 2014). The goal of those centers is to reduce the energy demand of the increasing number of servers required both to store and to process the big amount of data. These centers use technologies such as virtualization software that decreases the number of servers needed for operations and technologies to cut energy consumption and clean renewable power.

In Goiri et al. (2011, 2012), 'Parasol', a solar-powered micro data center, was developed backed by grid tie and batteries. Parasol can smartly schedule jobs and assign the energy source to be used focusing on the data center level design. In Sharma et al. (2011), 'Blink' was proposed which exploits internet workloads to leverage fast power state switching and match the server power demand with the available power budget. The Net-zero system designed by Banerjee et al. (2012) uses solar powered racks to match the load energy consumption with the renewable energy power supply based on grid-dependent power synchronization mechanism. In Li et al. (2015), 'iSwitch' was developed as renewable energy power tuning scheme capable of managing intermittent renewable power while governing the required performance levels and maximizing energy utilization with minimizing load matching activities in the data center. In Liu et al. (2018), a workload management system was proposed to maximize the efficiency in data centers by integrating the use of IT to measure the demand shifts to exploit time variations in electricity price, renewable energy generation rate and cooling efficiency. Another project, 'Oasis' was developed by Li et al. (2013) to scale out server clusters using incremental renewable energy integration at the protocol data unit level to add server racks to the existing data center.

In Shyam et al. (2015), 'Apache spark' was used as a cluster computing platform for smart grids. It combines batch and real-time data processing techniques with utilizing machine learning algorithms. In Baker et al. (2015), the energy efficiency of cloud routing was addressed rather than data centers energy consumption. The authors proposed 'GreeDi' a network-based energy efficient routing framework for storing and processing big data. The proposed algorithm was formalized by situation calculus, linear, goal and dynamic programming models. Recently, in Suryadevara (2021), an efficient machine learning algorithm was developed that works as a classifier to reduce energy levels with using sensors in an IoT application.

The research gap is to develop an approach that can identify the stability of a grid system at a low cost; a predictive model is needed to identify the stability of a complex grid system.

3. Main technologies of big data in the energy sector

The U.S. Department of Energy (2009) has identified six objectives of developing smart grids:

(1) Enabling customers to have effective participation.
(2) Accommodating to all options of generation and storage.
(3) Offering new products, services and markets.
(4) Providing an adequate level of power quality that can meet wide range of needs.
(5) Optimizing asset utilization and increasing efficiency.
(6) Enabling quick response to disturbances and emergencies.

Big data techniques can be employed to serve these objectives. This technology does not conflict with traditional pre-processing methods; instead, it is a combination between block calculations and traditional clustering to develop comparative analysis (He et al., 2017). According to Jiang et al. (2016), there are four main categories of big data key technologies used in the energy sector: Data acquisition and storing,

Data correlation analysis, Crowd-sourced data control and Data visualization; detailed description of these main technologies was illustrated in Hu et al. (2014). In the essence of renewable energy grids, knowledge moves from measurement collection and data conversion to information, then using the extracted information to expand knowledge, and finally the accumulation of this knowledge provides the wisdom required for a decision-maker.

There are two main types of renewable energy data: geospatial and temporal data. Geospatial data is concerned with the locations, while temporal data is concerned with data time characteristics. For renewable energy, Geospatial data may include the location of transmission infrastructure, cities, factories, hospitals, schools, roads, etc. (Shekhar et al., 2012); this data is based mainly on Geographical Information Systems (GIS) tools. Temporal data may include the consumption patterns with respect to time (annually, monthly, weekly, daily, and hourly) besides the amount of energy (e.g., sunshine) during different times of day or year. For small regions, such data can be available through traditional IT systems (such as SCADA). In the era of Internet of Things (IoT), smart buildings can provide their own data by means of smart metering and sensors (Mostafa et al., 2019; Ren et al., 2021). According to Niemi et al. (2012), there exist some simplified methods to calculate this data in case of lacking accurate measures or sensors. Most likely, the energy consumption peaks towards the urban center and decays towards the outskirts (Willis & Northcote-Green, 1983). Given a radial profile r that can be transformed into Cartesian coordinates (x, y), then r2 = x2 + y2. If the city center expressed by ($r = 0$), the radial load density (expressed in Megawatts/km2) peaks at point rm, the load profile expression will be:

$$Q^{1}(r,t) = \sum_{m=1}^{n} Q_{m}^{l}(r)\,\beta_{m}^{l}(t) = \sum_{m=1}^{n} P_{m}^{l} e^{-\alpha_{m}(r-r_{m})^{2}} \beta_{m}(t) \quad (1)$$

Where Q is the spatial load, l refers to the energy type, m denotes the load component, P is the maximum power density of a specific load, α is a width parameter, and β is a normalization function of load component time variation. To calculate the total load Equation (1) is integrated over area and time intervals.

A third type user classification data can be the social classification, users can be classified into categories not only according to geographic areas but also to their social stratums that can be an indicator for daily consumption curves (Zhou et al., 2016a). The weather data (e.g., angle of the sun rays, wind speed and direction, temperature, pressure, cloud cover, humidity, etc.) play a basic role in decision-making support in power stations (Zhou et al., 2016b). Hence, the integration between supply and demand data, spatial data, and temporal data can support strategic decisions such as location selection for renewable energy stations to improve output, productivity and efficiency. For a comprehensive review on big data and its techniques for energy systems, the reader is referred to the works by Jiang et al. (2016), Molina-Solana et al. (2017), Ma et al. (2017).

4. Big data framework for renewable energy grids

According to the literature, data analysis and decision making can be supported via a massive amount of data that should be stored and processed timely. The data includes: consumption rates, utilization patterns, waves' synchronization, maintenance schedules and reports, financial data, etc. Modern advances of communications made it possible to deliver real-time data and to manage the demand/supply equilibrium. Sometimes, traditional IT systems cannot detect the power system oscillations. Such applications are significant when using renewable energy sources, as these sources are most likely to cause unpredictable stress on the long-distance transmission lines connecting the grid with remote areas. Adequate management of big data can facilitate the demand response in power grids, electric vehicles and distributed energy resources (Bhattarai et al., 2019; Wang et al., 2019). Hence, big data can provide better and more secured bidirectional communication between different points to promote the energy resources in the energy markets.

The forecasting of future needs of a power system in general and smart grids particularly is considered as another aspect. Power utilities use BDA to estimate several parameters that support decision-making processes, such as load planning and power commitment (Rahman et al., 2016). This is especially important for renewable energy sources to estimate the available energy and the ability of the grid to transmit it. Other important applications include calculating the equipment downtime and estimating and analysis system failures. Therefore, improving the efficiency and robustness of the generation and distribution functions can be performed. Several tools were developed for big data streaming and operation in power systems. Some examples are Hadoop (Karun & Chitharanjan, 2013), Apache Drill (Chandarana & Vijayalakshmi, 2014), and Storm (Maske & Prasad, 2015). Through these technologies, users can have real-time interaction with machines. Following the layered structure developed by Hu et al. (2014), a big data framework for renewable energy power utility is developed in Fig. 1.

From Fig. 1, the big data framework consists of three layers. The upper and less complex layer is dedicated to storing data, accessing data and doing computations. The middle layer is responsible for managing and sharing data, integrate data between different applications and areas; data privacy is a key issue in this layer. In the bottom and deepest layer, the data mining platform is used for data preprocessing through data fusion technology. Due to the increasing volume of data, data storage and management are divided into a number of subtasks that are executed on several computing nodes. This big data structure is used to perform several functions such as failure event analysis (Qiu et al., 2018), risk analysis (Kezunovic et al., 2018), forecasting (Haupt & Kosovic, 2016), maintenance management (Zhou & Yang, 2016), and asset condition monitoring and assessment (Liu et al., 2016), evaluating the quality of Heating, Ventilation, and Air Conditioning (HVAC) systems in buildings (Barbeito et al., 2017).

To serve the framework depicted in Fig. 1, an integrated architecture based on BDA and cloud computing can be proposed. The key parts of this architecture are the smart grid, big data tool, database and the cloud environment. The big data tool is used for managing the storage and retrieval of data and the distributed storage to the nodes in racks (Pal & Agrawal, 2014). The database stores data about customer consumption patterns, historic data on supply, demand, failures, etc. Prediction algorithms are used to estimate the demand and supply of the grid (Wang et al., 2018). In Singh and Yassine (2018), the smart meter data analytics was addressed and proposed three main applications: load analysis, load forecasting and load management. The key techniques for these applications include time series, data clustering, dimensionality reduction, data classification, outlier detection, low-rank matrix and online learning. In Kezunovic et al. (2013), a smart data mining model was proposed that can be used for analyzing, predicting, and visualizing energy time series consumption patterns. A key part of this model was using 'Frequent pattern mining'; frequent patterns are item sets that appear in a dataset with frequency equal or more than a user-specified threshold. Frequent pattern mining is an essential data mining tool to process and analyze big data.

The ultimate objective of the described framework is to support decision-making by providing the information that can help to explore innovative solutions and applications. In Lin et al. (2012), four different types of decisions were identified based on the data analysis in power systems as listed in Table 2.

Data security is a key issue in BDA (Kezunovic et al., 2013), therefore it is of paramount importance to secure the distributed energy routing process against possible false data attacks. Different types of attacks may occur to manipulate energy supply, energy response or the link state of energy transmission (Lin et al., 2012). Such false data can lead to supply/demand imbalance and cause overhead costs and energy

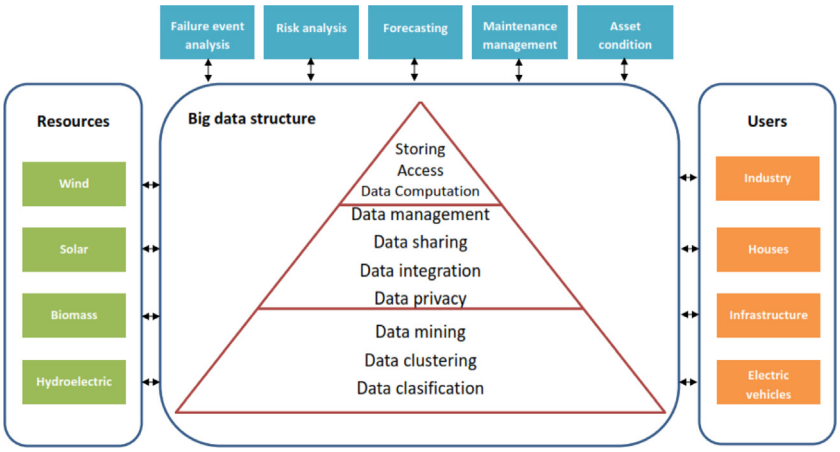

Fig. 1. A big data framework for renewable energy power utility.

Table 2
Types of decisions in a power system.

Decision type	Time frame	Objective	Example
Corrective	Just-in-time	Immediate handling of undesirable situations.	• Maintenance optimization • Risk-based asset management
Distributed	Daily	Assessment of system state.	• Online assessment of voltage stability
Adaptive	Daily	Monitoring events and operations.	• Online outage management • Disturbance detection.
Predictive	Weekly	Detailed forecasting of system behavior.	• Operations planning convergence • Interactions of renewable generation and loads

shortage. Assume a grid with a set of N customers. Let D_i^T denotes true demand by a customer $I \epsilon N$ and D_i^F denotes false demand for the same customer injected to the system by some hacker or competitor, and then a false energy-request messages will be sent to the energy-demand nodes in the grid. When the grid has the capacity to meet this demand, then that customer will receive more energy than what he truly requested. This leads to a loss in the quantity of supplied energy to customer I; this loss can be expressed by:

$$\Delta D_i = D_i^F - D_i^T \tag{2}$$

If the demand of several customers was hacked, then the total loss in the quantity of supplied energy in the grid is expressed by:

$$\Delta D = \sum_{i \in N^F} \Delta D_I \tag{3}$$

Where N^F is the set of customers with false demand. Equations (2) and (3) assume that the grid supply can fulfill all the requested demand even with the extra false amount. But what if the grid supply capacity cannot provide such amount? In that case, some customers of the grid will suffer from power shortage.

5. Predicting the smart grid stability through big data analytics

The smart grid system requires information about the consumer demand and the amount of the supplied energy, as well as the estimated grid stability to create new pricing for each energy unit sustainability of the smart grid system. Our research aims to predict and analyze the changes in energy production and consumption relative to the energy prices in a decentralized smart grid system, by performing BDA for a large dataset. To identify the grid stability for the distributed smart grid system, a mathematical model was presented in (Schäfer et al., 2016) for a four-node star architecture, with one energy source supplying three consumption nodes as displayed in Fig. 2, the model considers three input features; total power balance, energy price elasticity, and response time to price changes.

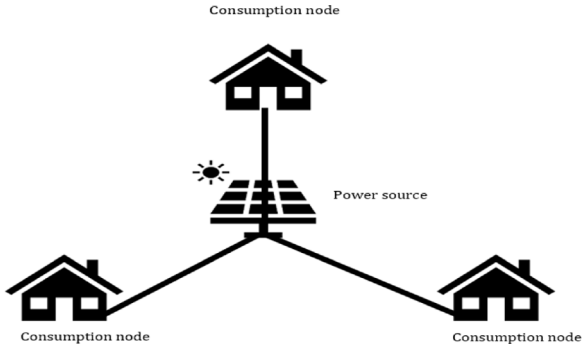

Fig. 2. Descriptive diagram for a 4-node star architecture in a smart grid system.

The decentralized smart grid data system was gathered by Arzamasov et al. (2018) is used, the dataset consists of 60,000 instances and 12 attributes, and the objective is to predict the stability of the decentralized smart grid control system through machine learning and deep learning. The data contains information about the demand input and grid output of the decentralized smart grid control system collected from various resources. The steps of predicting the stability of the smart grid system by using BDA is depicted in Fig. 3.

5.1. Creating the cloud storage and choosing the cloud computing platform

Cloud storage and cloud computing are considered critical steps for BDA and computations, the cloud storage platforms available for BDA are Amazon S3, BigQuery, google drive, Microsoft Azure, and Hadoop. Google Colab and Google drive (Bisong, 2019) are used in our case study as cloud computing platform for predicting the smart grid stability by using the Python coding (Van Rossum & Drake, 2009) and Apache Spark (Zaharia et al., 2016). The BDA was implemented in Python 3.0 on Google Collaboratory and Pyspark 3.1.2 was used

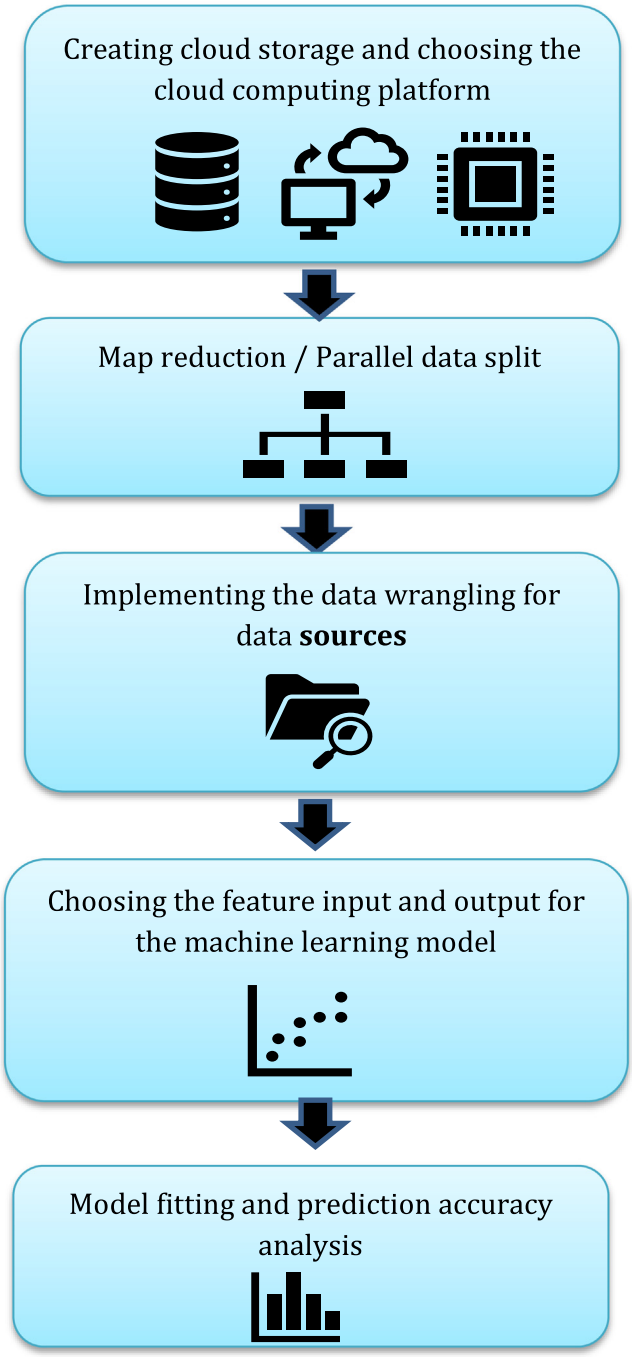

Fig. 3. Steps of predicting the smart grid stability for BDA.

to create the data pipeline. The computations were performed on a Lenovo Intel® Core I i5 2.70 GHz with 12 GB RAM running Windows 10 Professional operating system.

5.2. Map reduction/parallel data split

Map reduction data split is considered as essential step while handling big data. Map reduction aims to reduce the amount of data to be processed for parallel computing and the parallel data split is used, as well, to divide the amount of data into homogeneous sections for enhancing the speed of data processing. In our case study, the Apache Spark was successful in splitting the data and assessing the prediction accuracy for each data segment by using a training penalized regression model.

Table 3
Summary of the features dataset to the smart gird system big data analytics.

Summary	Count	Min	Max
tau1	60,000	0.5	10
tau2	60,000	0.5	10
tau3	60,000	0.5	10
tau4	60,000	0.5	10
p1	60,000	1.5	6
p2	60,000	−2	−0.5
p3	60,000	−2	−0.5
p4	60,000	−2	−0.5
g1	60,000	0.05	1
g2	60,000	0.05	1
g3	60,000	0.05	1
g4	60,000	0.05	1
Stab	60,000	−0.08	0.11
Stabf	60,000	Unstable	Stable

5.3. Implementing the data wrangling for data sources

The classification model was developed by using supervised machine learning, the output of the stability columns need to be defined as an integer instead of using a character or string, thus the data wrangling was essential for big data. Data wrangling can be applied simply using either a Pandas data frame or Spark SQL in the Apache Spark libraries. Some exploration was conducted through the histogram plotting for the stability values and stability classification of various connections in the smart grid network as shown in Fig. 4, and the pie chart that describes the percentage of stability, shown in Fig. 5.

The model contains both discrete and continuous, where the stability was scored by a stability index under ‘stab’ column (Continuous output) containing positive and negative values, and thus regression is used. Also, another output was provided under ‘stabf’ column (Discrete output); if stability index is positive it is labeled ‘stable’, and on the other hand if the stability index is negative, it is labeled ‘unstable’.

5.4. Choosing the feature input and output for the machine learning model

To identify the stability of the smart grid system, the dataset was gathered from the installation of a smart grid system in Karlsruhe, Germany. The list of inputs and outputs for the dataset were identified as follows:

5.4.1. Input features

- ‘tau1’ to ‘tau4’: the reaction time of each network participant, a real value within the range 0.5 to 10 (‘tau1’ corresponds to the supplier node, ‘tau2’ to ‘tau4’ to the consumer nodes).
- ‘p1’ to ‘p4’: nominal power produced (positive) or consumed (negative) by each network participant, a real value within the range −2.0 to −0.5 for consumers (‘p2’ to ‘p4’). As the total power consumed equals the total power generated, p1 (supplier node) = - (p2 + p3 + p4).
- ‘g1’ to ‘g4’: price elasticity coefficient for each network participant, a real value within the range 0.05 to 1.00 (‘g1’ corresponds to the supplier node, ‘g2’ to ‘g4’ to the consumer nodes; ‘g’ stands for ‘gamma’).

A description of the input and output features is shown in Table 3, containing all the discrete and continuous outputs used for regression and classifications

The choice of features was applied by using feature engineering, where each of the input features was tested by using *p*-value for their significance towards the labeled output, the features with significant p-values were selected. The results of the *p*-value hypothesis test showed

Fig. 4. Histogram for the stability index values as part of the data exploration and wrangling.

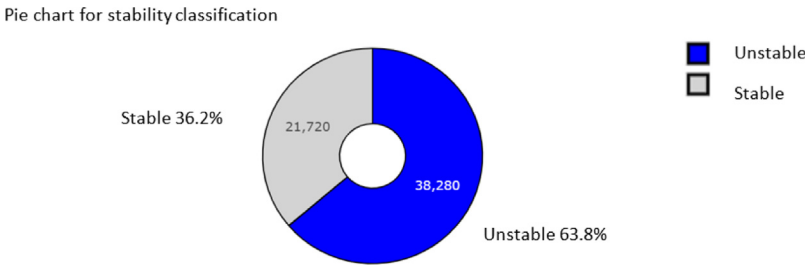

Fig. 5. Pie chart for the stability classification as part of the data exploration and wrangling.

Table 4
Application of feature engineering and selection of the input features for classification.

Features	P-value	Correlation	Significance
tau1	0.0001	−0.2181	Significant
tau2	0.0001	−0.2208	Significant
tau3	0.0001	−0.2249	Significant
tau4	0.0001	0.1884	Significant
p1	0.97	0.0024	Insignificant
p2	0.98	0.0016	Insignificant
p3	0.91	0.0007	Insignificant
p4	0.95	−0.0048	Insignificant
g1	0.0001	0.3920	Significant
g2	0.0001	−0.4202	Significant
g3	0.0001	−0.4184	Significant
g4	0.0001	−0.1822	Significant

the significance of the features related to the reaction time and the price elasticity for each network as well as the insignificance of the nominal power features as summarized in Table 4.

As part of the data exploration and analysis, the histogram of variables is used for providing a descriptive figure for the distribution of the dataset as shown in Fig. 6. Also, the correlation matrix in Fig. 7 is used to provide the correlation between the input feature with each other as well as the input feature with the output features, the strongest correlation appeared in the nominal power feature with each other despite of their insignificance to the output stability index and label.

5.4.2. Output of the machine learning model

- 'stab': the maximum real part of the characteristic differential equation root (if positive, the system is linearly unstable; if negative, the system is linearly stable); the 'stab' is used in the regression model due to the continuous output provided.
- 'stabf': a categorical (binary) label ('stable' or 'unstable'). The 'stabf' is used in the classification model as it contains discreate labeled output.

5.5. Model fitting and prediction accuracy analysis

Four machine learning models have been provided; the first model is related to the classification model, where the aim was to identify if the smart grid system will be considered stable or not, three machine learning models were provided using a decision tree, random forest classifier, and conventional neural network (Deep learning). The fourth machine learning model has used a penalized linear regression for the prediction of the stability differential equation root.

5.5.1. Decision tree algorithm

The decision tree is the usage of nodes and branches to enable the model to learn and identify the fitness accuracy, the difference between the decision tree and the random forest is the complexity of nodes and branching.

5.5.2. Random forest algorithm

Random forest algorithm is composed of different decision trees using the same node, the optimal solution is provided through the merging of various decision trees as illustrated in Fig. 8.

For classification problems, the Gini index is required to identify the nodes on the decision tree branch, this is expressed by Eq. (4).

$$\text{Gini} = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{4}$$

Where c is the number of classes and p_i is the relative frequency for each class.

5.5.3. Deep learning

Deep learning is a subfield of machine learning that utilizes the artificial neural network methodology to enhance model learning and data fitting (Gencer & Başçiftçi, 2021; Shariati et al., 2021). The equation for deep learning is given in Eq. (5) and is used to create the prediction model for the decentralized smart grid control system.

$$z = \sum_{i=1}^{12} w_i + x_i + b \tag{5}$$

Where z is the predicted output value from the deep learning model, i is the annotation for the variable, b is the bias, and w is the weight of each variable. The neural network is a multiple layer perception consisting of various layers with each layer contains various neurons as shown in Fig. 9, the layers are divided into:

- Input layers that represent the number of variables; 12 neurons.
- Output layer that represents the number of output classification; 2 neurons.

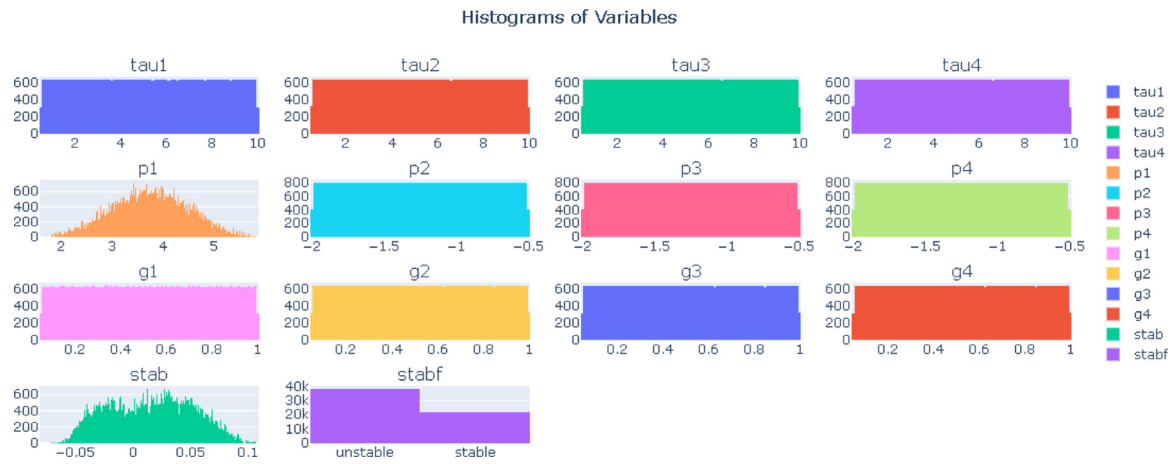

Fig. 6. Histogram for the input features as part of the data exploration and wrangling.

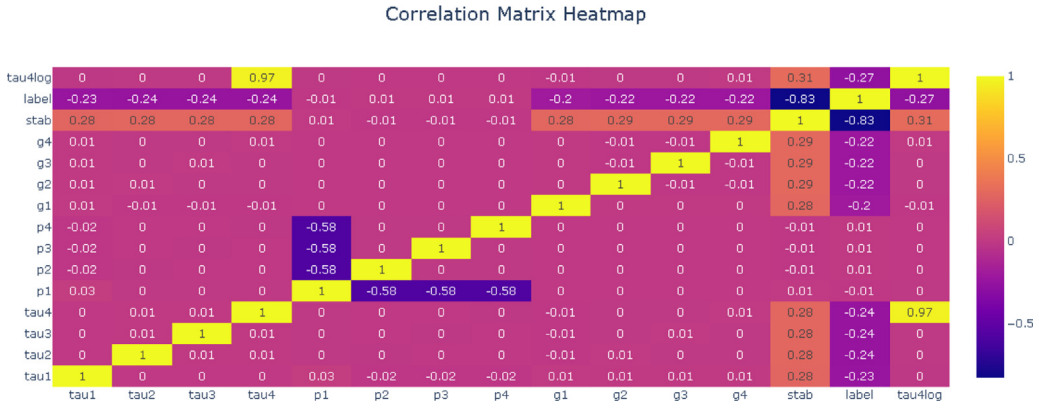

Fig. 7. Correlation matrix display as part of the data exploration and wrangling.

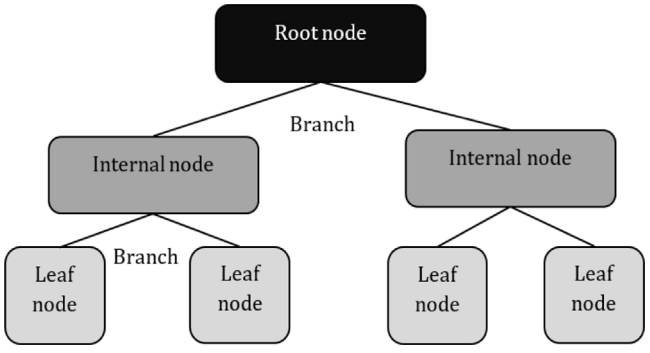

Fig. 8. Diagram showing the random forest algorithm structure used for smart grid BDA.

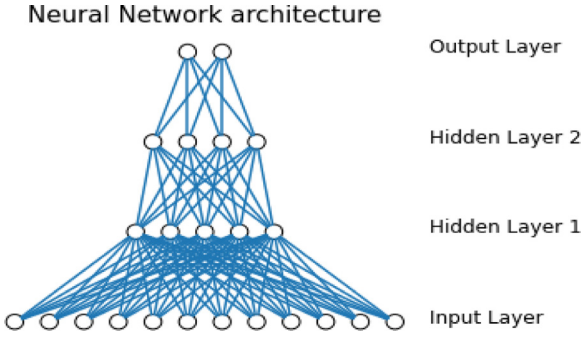

Fig. 9. The neural network structure of the deep learning model used for the smart grid BDA.

- Hidden layers that are intermediate layers used to model the decentralized smart grid system; the first layer contains 5 neurons, and the second layer contains 4 neurons.

To calculate the optimal weight for the neural network model, the model fitting must be performed through several iterations with the number of iterations specified as epochs (Smith, 2017). The activation function is the transfer function used to select the data used in learning propagation for the prediction model. The used function is the sigmoid function (Smith, 2017), as shown in Eq. (6).

$$\psi(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Where the $\psi(x)$ is the activation function used in the deep learning model developed.

5.5.4. Convolutional neural network

The convolution neural network is named after the usage of the mathematical linear operation between matrices called convolutional (Albawi et al., 2017). The convolutional neural network is used to get an abstract feature when the input feature propagates to a deeper layer in image processing, voice and text recognition. The convolutional neural network was applied for big data analytics from healthcare communities by Chen et al. (2017) and applied for big data analytics from

Indian customer sentiment using text on social applications regarding the electrical vehicles satisfaction as described by Jena (2020). The rule of convolutional formula is describe as shown in Eq. (7)

$$net(i,j) = (x * w)[i,j] = \sum_m \sum_n x[m-n]w[i-m, j-n] \quad (7)$$

Where $net(i,j)$ is the output to the next layer, the x is the input text to the layer, w is the kernel filter, and x is the convolutional operation.

5.5.5. Penalized linear regression

The penalized linear regression aims to reduce the mean square errors of the model by identifying the vectors for the real number β_0 and β as shown in Eq. (8) (Jozaghi et al., 2019).

$$\beta_0^*, \beta^* = argmin(\frac{1}{n}\sum_{i=1}^{n}(y_i - (x_i * \beta + \beta_0))^2) \quad (8)$$

Where y_i is the predicted output, n is the number of attributes, x_i is the input variable.

5.5.6. Gradient boost decision tree method

The gradient boost is used for classification problems to obtain an accurate predictive classification model by combining different base classifiers into a strong base classifier as stated by Li et al. (2020). The gradient boost decision tree provides a global convergence for the algorithm towards the negative gradient for the predictive loss function. Henceforth, it can provide high accuracy for the predictive classification model. The equation of the gradient boost decision tree method is described in Eq. (9)

$$\alpha_m, \beta_m = argmin(\sum_{i=1}^{n}(L(Y_i) - F_m(x_i)) + \beta h(x_i; \propto)) \quad (9)$$

Where α_m, β_m are optimal parameters for the loss function, $F_m(x_i)$ is the prediction function obtained for the mth iteration, $h(x_i; \propto)$ is a base learner model developed as a simple function of x_i and parameter $\propto$

6. Results and discussion

The results of the penalized learning model are shown in Fig. 10. Fig. 11(a and b) gives a comparison between the classification algorithms used for BDA of the smart grid dataset in terms of accuracy and running time, respectively. The results show an accuracy of 96% for the regression model implemented using 70% of the data as a training set, and 30% as the testing set. Using the random forest tree model yielded 84% accuracy, the decision tree model yielded 78% accuracy, and the convolutional neural network (CNN) yielded 87% for classification model, and the gradient boosted decision tree model resulted in 87%. Thus, the regression model has shown the highest accuracy in predicting the system stability through numerical values as shown in Fig. 10 and the convolutional neural network was able to provide the highest accuracy in predicting the smart grid stability through classification model at a faster time compared to other machine learning classification models as shown in Fig. 11. The predictive model produced from either CNN or regression model could be used in predicting the stability of more complicated smart grid systems across the EU and optimizing the smart grid setting of various input and output power sources connection to provide the maximum stability. The amount of data available in the dataset is considered relatively small for big data analytics, however the cloud computing provided was suitable for big data analytics framework.

The improvement of smart grid stability can be provided from using the CNN model and setting up the optimal settings from linkage between different power source and consumption outlets. The results have shown an optimized model accuracy and fast processing for the CNN compared to other ML classification algorithms.

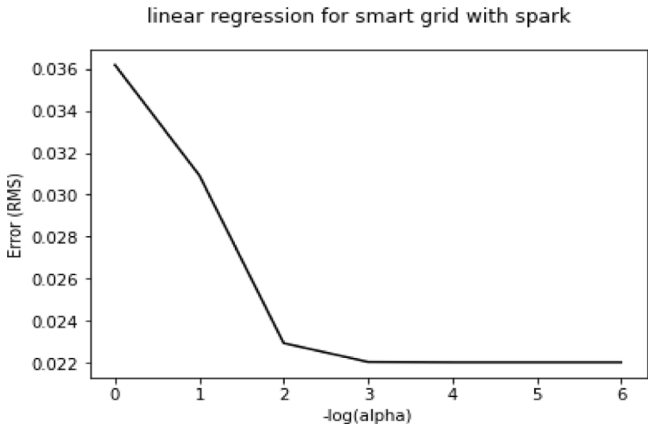

Fig. 10. Penalized linear regression graph for the smart grid stability.

7. Conclusions and recommendations

Shifting towards sustainable energy systems requires advanced technologies such as smart grids, renewable energy storage and management power stations. Managing and operating such complex systems depends on integrating and coordinating several components. Advanced sensors and meters are used as main data sources in the era of big data, wireless communication, and IoT. The data captured by such tools should be stored, processed and analyzed to get the necessary information for the deployment of smart grids and the efficient operation of power stations demand and supply with integrating renewable energy. Improving fuel efficiency besides minimizing emissions and waste should be also accounted for. Many benefits can be obtained from implementing such technology in the areas of asset management, operations planning, monitoring voltage instability, stability margin prediction and fault detection. Accordingly, many challenges are associated with such implementation; the main issues are related to data uncertainty, data quality, data security, and data complexity.

In this work, a big data framework was built to identify the stability of the smart grid dataset consisting of 60,000 instances and 12 attributes. The BDA framework was coded and implemented in Python on Google Collaboratory, and Pyspark was used to create the data pipeline. For larger datasets, it is recommended to use Amazon S3 storage and EC2 for cloud computing. The results have shown the high accuracy of the penalized linear regression for fitting a regression model on decentralized smart grid control system with BDA as well as the high accuracy and fast computation of neural networks in fitting a classification model for the decentralized smart grid system in comparison with other classification models such as random forest and decision tree.

This paper has three main contributions; first, it provides an exhaustive review on the previous recent works that have addressed the use of BDA in energy applications. The second contribution is exhibiting a framework of the potential implementation of BDA for smart grids and renewable energy power utilities. The third contribution is proposing a five-step approach for predicting the smart grid stability for BDA by using three different machine learning methods. The predictive model could be used later for optimizing the smart grid setting of various input and output power sources. Thus, enhancing the smart grid stability

Limitations of the research is that even though the 60,000 objects do not represent big data, cloud computing and cloud storage provided using the 60,000 objects has been useful in providing a big data framework, Also the real time event analysis for the data gathered was simulated by using Apache spark and Google Colab.

Future research should address these different crucial recommendations and perspectives:

- Involvement of customers; through enabling creative solutions for customers to participate in data entry and providing data about

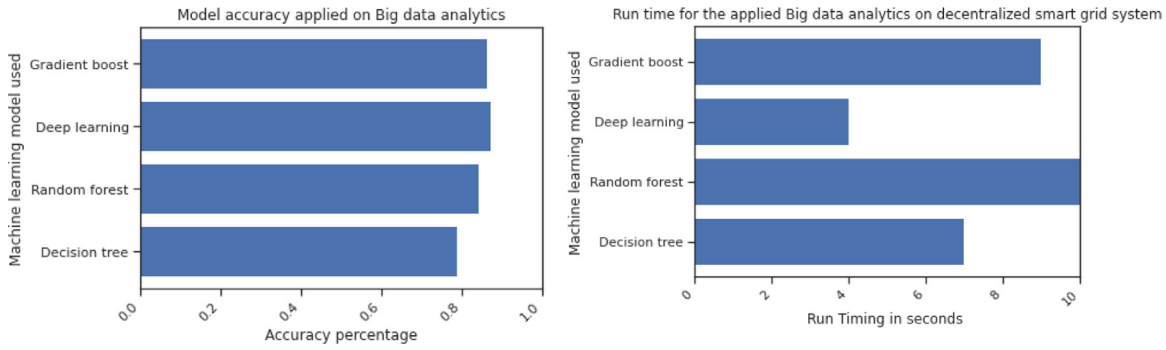

Fig. 11. Comparison between the classification algorithms used for BDA of the smart grid in terms of (a) accuracy (b) running time.

their demand and storage capacities. Therefore, more convenient and cost-efficient strategies can be offered and studied.

- Incorporation of integrated and seamless monitoring and control systems; through BDA tools and techniques to improve visualization and advanced real-time control and minimize risk.
- Modifications towards new regulatory environment for smart grid operations. The traditional economic principles should be changed to allow customers to modify their demand timings to respond to incentives. Hence, new demand response capabilities can be offered.
- Comparison between penalized linear regression and Support vector regression
- Collection of more data is needed to analyze the smart grid stability with variety of renewable energy sources and demand across more countries with the established big data analytics framework and cloud computing through Apache Spark and Google Colab.
- Collection of a larger dataset size with real time event analysis in order to fully utilize the big data analytics and verify the framework feasibility.
- Comparison with related works and explain the similarities and differences between different approaches.

CRediT authorship contribution statement

Noha Mostafa: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Haitham Saad Mohamed Ramadan:** Conceptualization. **Omar Elfarouk:** Software, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2022.100363.

References

Al-Ali, A. R., & Aburukba, R. (2015). Role of internet of things in the smart grid technology. *Journal of Computer and Communications*, *3*(5), 229–233. http://dx.doi.org/10.4236/jcc.2015.35029.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *Paper presented at the 2017 international conference on engineering and technology* (pp. 1–6). http://dx.doi.org/10.1109/ICEngTechnol.2017.8308186.

Aquino-Lugo, A. A., Klump, R., & Overbye, T. J. (2011). A control framework for the smart grid for voltage support using agent-based technologies. *IEEE T Smart Grid*, *2*(1), 173–180. http://dx.doi.org/10.1109/TSG.2010.2096238.

Arenas-Martinez, M., Herrero-Lopez, S., Sanchez, A., Williams, J. R., Roth, P., Hofmann, P., & Zeier, A. (2010). A comparative study of data storage and processing architectures for the smart grid. In *Proc IEEE int conf smart grid commun, Vol. 28* (pp. 5–290). http://dx.doi.org/10.1109/SMARTGRID.2010.5622058.

Arzamasov, V., Böhm, K., & Jochem, P. (2018). Towards concise models of grid stability. In *Paper presented at the 2018 IEEE international conference on communications, control, and computing technologies for smart grids* (pp. 1–6).

Asad, Z., & Chaudhry, M. A. R. (2017). A two-way street: Green big data processing for a greener smart grid. *IEEE Systems Journal*, *11*(2), 784–795. http://dx.doi.org/10.1109/JSYST.2015.2498639.

Avgerinou, M., Bertoldi, P., & Castellazzi, L. (2017). Trends in data centre energy consumption under the European code of conduct for data centre energy efficiency. *Energies*, *10*, Article 1470. http://dx.doi.org/10.3390/en10101470.

Baker, T., Al-Dawsari, B., Tawfik, H., Reid, D., & Ngoko, Y. (2015). GreeDi: An energy efficient routing algorithm for big data on cloud. *Ad Hoc Networks*, *35*, 83–96. http://dx.doi.org/10.1016/j.adhoc.2015.06.008.

Banerjee, P., Patel, C., Bash, C., Shah, A., & Arlitt, M. (2012). Towards a net-zero data center. *ACM Journal on Emerging Technologies in Computing Systems*, *8*(4), Article 27. http://dx.doi.org/10.1145/2367736.2367738.

Barbeito, I., Zaragoza, S., Tarrío-Saavedra, J., & Naya, S. (2017). Assessing thermal comfort and energy efficiency in buildings by statistical quality control for autocorrelated data. *Applied Energy*, *190*(C), 1–17. http://dx.doi.org/10.1016/j.apenergy.2016.12.100.

Berral, J., Gojri, I., Nguyen, T., Gavalda, R., Torres, J., & Bianchini, R. (2014). Building green cloud services at low cost. In *IEEE international conference on distributed computing systems, Vol. 44* (pp. 9–460). http://dx.doi.org/10.1109/ICDCS.2014.53.

Bhattarai, B. P. (2019). Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid*, http://dx.doi.org/10.1049/iet-stg.2018.0261.

Billinton, R., & Gao, Y. (2008). Multistate wind energy conversion system models for adequacy assessment of generating systems incorporating wind energy. *IEEE Transactions on Energy Conversion*, *23*(1), 163–170. http://dx.doi.org/10.1109/TEC.2006.882415.

Bisong, E. (2019). Google colaboratory. In *Building machine learning and deep learning models on google cloud platform*. Berkeley, CA: A Press, http://dx.doi.org/10.1007/978-1-4842-4470-8_7.

Chandarana, P., & Vijayalakshmi, M. (2014). Big data analytics frameworks. In *Proc IEEE int conf circuits, syst, commun inf technol appl, Vol. 43* (pp. 0–434). http://dx.doi.org/10.1109/CSCITA.2014.6839299.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, *5*, 8869–8879. http://dx.doi.org/10.1109/ACCESS.2017.2694446.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. http://dx.doi.org/10.1007/s11036-013-0489-0.

Diamantoulakis, P. D., Kapinas, V. M., & Karagiannidis, G. (2015). Big data analytics for dynamic energy management in smart grids. *Big Data Research*, *2*(3), 94–101. http://dx.doi.org/10.1016/j.bdr.2015.03.003.

Gencer, K., & Başçiftçi, F. (2021). Time series forecast modeling of vulnerabilities in the android operating system using ARIMA and deep learning methods. *Sustainable Computing: Informatics and Systems*, *30*, Article 100515. http://dx.doi.org/10.1016/j.suscom.2021.100515.

Goiri, I., Beauchea, R., Le, K., Nguyen, T., Haque, M., Guitart, J., Torres, J., & Bianchini, R. (2011). GreenSlot: Scheduling energy consumption in green datacenters. In *Proc of the int conf for high performance computing, networking, storage and analysis* (pp. 1–11). http://dx.doi.org/10.1145/2063384.2063411.

Goiri, I., Le, K., Nguyen, T., Guitart, J., Torres, J., & Bianchini, R. (2012). GreenHadoop: Leveraging green energy in data-processing frameworks. (pp. 7–70). http://dx.doi.org/10.1145/2168836.2168843.

Goyena, S. G., Sádaba, Ó. A., & Acciona, S. A. (2009). Sizing and analysis of big scale and isolated electric systems based on renewable sources with energy storage. In *IEEE PES/IAS conference on sustainable alternative energy* (pp. 1–7). http://dx.doi.org/10.1109/SAE.2009.5534837.

Haupt, S. E., & Kosovic, B. (2016). Variable generation power forecasting as a big data problem. *IEEE Transactions on Sustainable Energy*, *8*(2), 725–732. http://dx.doi.org/10.1109/TSTE.2016.2604679.

He, X., Ai, Q., Qiu, R. C., Huang, W., Piao, L., & Liu, H. (2017). A big data architecture design for smart grids based on random matrix theory. *IEEE Transactions on Smart Grids*, *8*(2), 674–686. http://dx.doi.org/10.1109/TSG.2015.2445828.

Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, *2*, 652–687. http://dx.doi.org/10.1109/ACCESS.2014.2332453.

Jena, R. (2020). An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach. *Industrial Marketing Management*, *90*, 605–616. http://dx.doi.org/10.1016/j.indmarman.2019.12.012.

Jiang, H., Wang, K., Wang, Y., Gao, M., & Zhang, Y. (2016). Energy big data: A survey. *IEEE Access*, *4*, 3844–3861. http://dx.doi.org/10.1109/ACCESS.2016.2580581.

Jozaghi, A., Shen, H., Ghazvinian, M., Seo, D. J., Zhang, Y., Welles, E., & Reed, S. (2019). Multi-model streamflow prediction using conditional bias-penalized multiple linear regression. *Stochastic Environmental Research and Risk Assessment*, *35*, 1–19. http://dx.doi.org/10.1007/s00477-021-02048-3.

Ju, X., Chen, V., Rosenberger, J., & Liu, F. (2021). Fast knot optimization for multivariate adaptive regression splines using hill climbing methods. *Expert Systems with Applications*, *171*, Article 114565. http://dx.doi.org/10.1016/j.eswa.2021.114565.

Ju, X., Liu, F., Wang, L., & Lee, W.-J. (2019). Wind farm layout optimization based on support vector regression guided genetic algorithm with consideration of participation among landowners. *Energy Conversion and Management*, *196*, 1267–1281. http://dx.doi.org/10.1016/j.enconman.2019.06.082.

Karun, A. K., & Chitharanjan, K. (2013). Locality sensitive hashing based incremental clustering for creating affinity groups in Hadoop-HDFS-an infrastructure extension. In *Proc. IEEE int. conf. circuits, power comput. technol, Vol. 124* (pp. 3–1249). http://dx.doi.org/10.1109/ICCPCT.2013.6528999.

Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., & Cunningham, R. K. (2014). Computing on masked data: a high-performance method for improving big data veracity. In *IEEE high performance extreme computing conference* (pp. 1–6). http://dx.doi.org/10.1109/HPEC.2014.7040946.

Kezunovic, M., Obradovic, Z., Djokic, T., & Roychoudhury, S. (2018). Systematic framework for integration of weather data into prediction models for the electric grid outage and asset management applications. In *Proceedings of the 51st hawaii international conference on system sciences, Vol. 273* (pp. 7–2746). http://dx.doi.org/10.24251/HICSS.2018.346.

Kezunovic, M., Xie, L., & Grijalva, S. (2013). The role of big data in improving power system operation and protection. IREP symposium Bulk power system dynamics and control - IX optimization. In *Security and control of the emerging power grid* (pp. 1–9). http://dx.doi.org/10.1109/IREP.2013.6629368.

Khalid, R., Javaid, N., Rahim, M. H., Aslam, S., & Sher, A. (2019). Fuzzy energy management controller and scheduler for smart homes. *Sustainable Computing: Informatics and Systems*, *21*, 103–118. http://dx.doi.org/10.1016/j.suscom.2018.11.010.

Koseleva, N., & Ropaite, G. (2017). Big data in building energy efficiency: understanding of big data and main challenges. *Procedia Engineering*, *172*, 544–549. http://dx.doi.org/10.1016/j.proeng.2017.02.064.

Kung, L., & Wang, H. F. (2015). A recommender system for the optimal combination of energy resources with cost–benefit analysis. In *International conference on industrial engineering and operations management* (pp. 1–10). http://dx.doi.org/10.1109/IEOM.2015.7093924.

Kwac, J., & Rajagopal, R. (2013). Demand response targeting using big data analytics. In *IEEE international conference on big data, Vol. 68* (pp. 3–690). http://dx.doi.org/10.1109/BigData.2013.6691643.

Labeeuw, W., & Deconinck, G. (2013). Residential electrical load model based on mixture model clustering and Markov models. *IEEE Transactions on Industrial Informatics*, *9*(3), 1561–1569. http://dx.doi.org/10.1109/TII.2013.2240309.

Lee, J., Bagheri, B., & Kao, H. A. (2014). Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. In *Proceedings of international conference on industrial informatics*. http://dx.doi.org/10.13140/2.1.1464.1920.

Li, C., Qouneh, A., & Li, T. (2012). Iswitch: Coordinating and optimizing renewable energy powered server clusters. In *In the 39th annual international symposium on computer architecture, Vol. 51* (pp. 2–523). http://dx.doi.org/10.1109/ISCA.2012.6237044.

Li, G., Wang, Y., He, J., Hao, Q., Yang, H., & Wei, J. (2020). Tool wear state recognition based on gradient boosting decision tree and hybrid classification RBM. *International Journal of Advanced Manufacturing Technology*, *110*(1–2), 511–522. http://dx.doi.org/10.1007/s00170-020-05890-x.

Li, C. (2013). Enabling datacenter servers to scale out economically and sustainably. In *Proceedings of the 46th annual IEEE/ACM international symposium on microarchitecture, Vol. 32* (pp. 2–333). http://dx.doi.org/10.1145/2540708.2540736.

Li, C. (2015). Towards sustainable in-situ server systems in the big data era. In *2015 ACM/IEEE 42nd annual international symposium on computer architecture, Vol. 1* (pp. 4–26). http://dx.doi.org/10.1145/2749469.2750381.

Lin, J., Yu, W., Yang, X., Xu, G., & Zhao, W. (2012). On false data injection attacks against distributed energy routing in smart grid. In *In Proceedings of the 2012 IEEE/ACM third international conference on cyber-physical systems, Vol. 18* (pp. 3–192). IEEE Computer Society Washington, http://dx.doi.org/10.1109/ICCPS.2012.26.

Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., Marwah, M., & Hyser, C. (2012). Renewable and cooling aware workload management for sustainable data centers. *ACM SIGMETRICS Performance Evaluation Review*, *40*(1), 175–186. http://dx.doi.org/10.1145/2318857.2254779.

Liu, H. (2016). The design and implementation of the enterprise level data platform and big data driven applications and analytics. In *IEEE/PES transmission and distribution conference and exposition*. http://dx.doi.org/10.1109/TDC.2016.7520032.

Liu, H. (2018). Thermal-aware and DVFS-enabled big data task scheduling for data centers. *IEEE Transactions on Big Data*, *4*, 177–190. http://dx.doi.org/10.1109/TBDATA.2017.2763612.

Luo, X., & Oyedele, L. O. (2022). A self-adaptive deep learning model for building electricity load prediction with moving horizon. *Machine Learning with Applications*, *7*, Article 100257. http://dx.doi.org/10.1016/j.mlwa.2022.100257.

Ma, Z., Xie, J., Li, H., Sun, Q., Si, Z., Zhang, J., & Guo, J. (2017). The role of data analysis in the development of intelligent energy networks. *IEEE Network*, *31*(5), 88–95. http://dx.doi.org/10.1109/MNET.2017.1600319.

Marmaras, C., Javed, A., Cipcigan, L., & Rana, O. (2017). Predicting the energy demand of buildings during triad peaks in GB. *Energy and Buildings*, *141*, 262–273. http://dx.doi.org/10.1016/j.enbuild.2017.02.046.

Mashayekhy, L., Nejad, M. M., Grosu, D., Zhang, Q., & Shi, W. (2014). Energy-aware scheduling of MapReduce jobs for big data applications. *IEEE Transactions on Parallel and Distributed Systems*, *26*(10), 2720–2733. http://dx.doi.org/10.1109/TPDS.2014.2358556.

Maske, M. M., & Prasad, P. (2015). A real time processing and streaming of wireless network data using storm. In *Proc of international conference on computation of power, energy, information and communication, Vol. 24* (pp. 4–249). http://dx.doi.org/10.1109/ICCPEIC.2015.7259471.

Mathew, P. A., Dunn, L. N., Sohn, M. D., Mercado, A., Custudio, C., & Walter, T. (2015). Big data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, *140*, 85–93. http://dx.doi.org/10.1016/j.apenergy.2014.11.042.

Missaoui, R., Joumaa, H., Ploix, S., & Bacha, S. (2014). Managing energy smart homes according to energy prices: Analysis of a building energy management system. *Energy and Buildings*, *71*, 155–167. http://dx.doi.org/10.1016/j.enbuild.2013.12.018.

Molina-Solana, M., Ros, M., Ruiz, M. D., Gómez-Romero, J., & Martin-Bautista, M. J. (2017). Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, *70*, 598–609. http://dx.doi.org/10.1016/j.rser.2016.11.132.

Mostafa, N., Hamdy, W., & Elawady, H. (2019). Impacts of internet of things on supply chains: A framework for warehousing. *Social Sciences: Industry 4.0 Implication for Economy and Society*, *8*(3), Article 84. http://dx.doi.org/10.3390/socsci8030084.

Mostafa, N., & Negm, A. (2018). Promoting organizational sustainability and innovation: An exploratory case study from the Egyptian chemical industry. *Procedia Manufacturing*, *22*, 1007–1014. http://dx.doi.org/10.1016/j.promfg.2018.03.143.

Niemi, R., Mikkola, J., & Lund, P. D. (2012). Urban energy systems with smart multi-carrier energy networks and renewable energy generation. *Renewable Energy*, *48*, 524–536. http://dx.doi.org/10.1016/j.renene.2012.05.017.

Pal, A., & Agrawal, S. (2014). An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using HDFS and MapReduce. In *Proceedings IEEE int. conf. netw. soft comput., Vol. 44* (pp. 2–447). http://dx.doi.org/10.1109/CNSC.2014.6906718.

Pan, E., Wang, D., & Han, Z. (2016). Analyzing big smart metering data towards differentiated user services: A sublinear approach. *IEEE Transactions on Big Data*, *2*, 249–261. http://dx.doi.org/10.1109/TBDATA.2016.2599924.

Qiu, R., Chu, L., He, X., Ling, Z., & Liu, H. (2018). Spatio-temporal big data analysis for smart grids based on random matrix theory: A comprehensive study. In HT Mouftah, M Erol-Kantarci, & M. H. Rehmani (Eds.), *Transportation and power grid in smart cities: communication networks and services*. Wiley, http://dx.doi.org/10.1002/9781119360124.ch23.

Rahimi-Eichi, H., & Chow, M. Y. (2014). Big-data framework for electric vehicle range estimation. In *The 40th annual conference of the IEEE industrial electronics society, Vol. 562* (pp. 8–5634). http://dx.doi.org/10.1109/IECON.2014.7049362.

Rahimi-Eichi, H., Jeon, P. B., Chow, M. Y., & Yeo, T. J. (2015). Incorporating big data analysis in speed profile classification for range estimation. In *IEEE 13th international conference on industrial informatics, Vol. 129* (pp. 0–1295). http://dx.doi.org/10.1109/INDIN.2015.7281921.

Rahman, M. N., Esmailpour, A., & Zhao, J. (2016). Machine learning with big data an efficient electricity generation forecasting system. *Big Data Research*, *5*, 9–15. http://dx.doi.org/10.1016/j.bdr.2016.02.002.

Rathor, S. K., & Saxena, D. (2020). Energy management system for smart grid: An overview and key issues. *International Journal of Energy Research*, *44*(6), 4067–4109. http://dx.doi.org/10.1002/er.4883.

Ren, G., Yu, M., Yin, D., Huang, S., Xu, H., & Yuan, M. (2021). Design and optimization of integrated energy management network system based on internet of things technology. *Sustainable Computing: Informatics and Systems*, *30*, Article 100502. http://dx.doi.org/10.1016/j.suscom.2020.100502.

Rifkin, J. (2011). *The third industrial revolution: how lateral power is transforming energy, the economy, and the world*. New York: St. Martin's Griffin.

Rogers, K. M., Klump, R., Khurana, H., Aquino-Lugo, A. A., & Overbye, T. J. (2010). An authenticated control framework for distributed voltage support on the smart grid. *IEEE Transactions on Smart Grids*, *1*(1), 40–47. http://dx.doi.org/10.1109/TSG.2010.2044816.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems, Vol. 4* (pp. 2–47). http://dx.doi.org/10.1109/CTS.2013.6567202.

Schäfer, B., Grabow, C., Auer, S., Kurths, J., Witthaut, D., & Timme, M. (2016). Taming instabilities in power grid networks by decentralized control. *The European Physical Journal Special Topics*, *225*(3), 569–582. http://dx.doi.org/10.1140/epjst/e2015-50136-y.

Shariati, M., Mafipour, M. S., & Mehrabi, P. (2021). A novel approach to predict shear strength of tilted angle connectors using artificial intelligence techniques. *Engineering with Computers*, *37*, 2089–2109. http://dx.doi.org/10.1007/s00366-019-00930.

Sharma, N., Barker, S., Irwin, D., & Shenoy, P. (2011). Blink: Managing server clusters on intermittent power. *Peer Journal Computer Science*, Article e34. http://dx.doi.org/10.7717/peerj-cs.34.

Shekhar, S., Gunturi, V., Evans, M. R., & Yang, K. (2012). Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the eleventh acm international workshop on data engineering for wireless and mobile access* (pp. 1–6). http://dx.doi.org/10.1145/2258056.2258058.

Shyam, R., Bharathi, G. H. B., Sachin, K. S., Prabaharan, P., & Soman, K. P. (2015). Apache spark a big data analytics platform for smart grid. *Procedia Technology*, *21*, 171–178. http://dx.doi.org/10.1016/j.protcy.2015.10.085.

Singh, S., & Yassine, A. (2018). Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies*, *11*, Article 452. http://dx.doi.org/10.3390/en11020452.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *IEEE winter conference on applications of computer vision, Vol. 46* (pp. 4–472). http://dx.doi.org/10.1109/WACV.2017.58.

Strohbach, M., Ziekow, H., Gazis, V., & Akiva, N. (2015). Towards a big data analytics framework for IoT and smart city applications. In F Xhafa, L Barolli, A Barolli, & P Papajorgji (Eds.), *Modeling and Processing for Next-Generation Big-Data Technologies*, *Modeling and optimization in science and technologies, Vol. 4* (pp. 257–282). Cham: Springer, http://dx.doi.org/10.1007/978-3-319-09177-8_11.

Su, W., & Chow, M. Y. (2012). Performance evaluation of an EDA-based large-scale plug-in hybrid electric vehicle charging algorithm. *IEEE Transactions on Smart Grids*, *3*(1), 308–315. http://dx.doi.org/10.1109/TSG.2011.2151888.

Suryadevara, N. K. (2021). Energy and latency reductions at the fog gateway using a machine learning classifier. *Sustainable Computing: Informatics and Systems*, *31*, Article 100582. http://dx.doi.org/10.1016/j.suscom.2021.100582.

Tannahill, B. K., & Jamshidi, M. (2014). System of systems and big data analytics – bridging the gap. *Computers and Electrical Engineering*, *40*(1), 2–15. http://dx.doi.org/10.1016/j.compeleceng.2013.11.016.

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, *11*(5), 239–273.

U. S. Department of Energy (2009). Smart grid system report. https://www.energy.gov/oe/downloads/2009-smart-grid-system-report-july-2009 (Accessed 5 March 2018).

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grids*, *10*(3), 3125–3148. http://dx.doi.org/10.1109/TSG.2018.2818167.

Wang, K., Xu, C., Zhang, Y., Guo, S., & Zomaya, A. (2019). Robust big data analytics for electricity price forecasting in the smart grid. *IEEE Transactions on Big Data*, *5*(1), 34–45. http://dx.doi.org/10.1109/TBDATA.2017.2723563.

Willis, H. L., & Northcote-Green, J. E. D. (1983). Spatial electric load forecasting: a tutorial review. *Proceedings of the IEEE*, *71*(2), 232. http://dx.doi.org/10.1109/PROC.1983.12562, e53.

Yan, Y., Qian, Y., Sharif, H., & Tipper, D. (2013). A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *IEEE Communication Surveys & Tutorials*, *15*(1), 5–20. http://dx.doi.org/10.1109/SURV.2012.021312.00034.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., & Dave, A. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65. http://dx.doi.org/10.1145/2934664.

Zhang, G., Bai, X., & Wang, Y. (2021). Short-time multi-energy load forecasting method based on CNN-Seq2Seq model with attention mechanism. *Machine Learning with Applications*, *5*, Article 100064. http://dx.doi.org/10.1016/j.mlwa.2021.100064.

Zhou, K., Fu, C., & Yang, S. (2016a). Big data driven smart energy management: From big data to big insights. *Renewable and Sustaianble Energy Reviews*, *56*, 215–225. http://dx.doi.org/10.1016/j.rser.2015.11.050.

Zhou, K., & Yang, S. (2016). Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustaianble Energy Reviews*, *56*, 810–819. http://dx.doi.org/10.1016/j.rser.2015.12.001.

Zhou, K., Yang, S., & Shao, Z. (2016b). Energy internet: The business perspective. *Applied Energy*, *178*, 212–222. http://dx.doi.org/10.1016/j.apenergy.2016.06.052.