IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 26, 2022, accepted February 25, 2022, date of publication March 8, 2022, date of current version March 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157941

A Survey of Preprocessing Methods Used for Analysis of Big Data Originated From Smart Grids

TURKI ALI ALGHAMDI[1] AND NADEEM JAVAID[2,3], (Senior Member, IEEE)

[1]Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[2]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad 44000, Pakistan
[3]School of Computer Science, University of Technology Sydney (UTS), Ultimo, NSW 2007, Australia

Corresponding author: Nadeem Javaid (nadeemjavaidqau@gmail.com)

ABSTRACT In this paper, a brief survey of data preprocessing methods is presented. Specifically, the data preprocessing methods used in the smart grid (SG) domain are surveyed. Also, with the advent of SG, data collection on a large scale became possible. The data is essential for electricity demand, generation and price forecasting, which plays an important role in making energy efficient decisions, and long and short term predictions regarding energy generation, consumption and storage. However, the forecasting accuracy decreases when data is used in raw form. Hence, data preprocessing is considered essential. This paper provides an overview of the data preprocessing methods and a detailed discussion of the methods used in the existing literature. A comparison of the methods is also given. A survey of closely related survey papers is also presented and the papers are compared based on their contributions. Moreover, based on the discussion of the data preprocessing methods, a narrative is built with a critical analysis. Finally, future research directions are discussed to guide the readers.

INDEX TERMS Data analytics, data preprocessing, integration, normalization, smart grid, smart meter, transformation.

I. INTRODUCTION

Today, due to the latest innovations and advancements in the technological sector, the energy demand has increased manifolds. Besides, energy demand is also influenced by population increase [1]. To satisfy the energy demand and supply needs, the traditional grid is the major source of energy generation [2]. However, because of the government policies, high electricity generation and consumption costs, and other factors, the traditional grid cannot sufficiently meet electricity demand. The concept of a smart grid (SG) is introduced to address these problems [3].

Information and communication technologies are integrated into the traditional grids for enabling bidirectional communication between the utility and the customers. Therefore, it is easy to capture a huge amount of electricity consumption data from smart meters [4]. SGs can efficiently sense the data present at the demand and supply sides. Real time data is captured from various points using sensors and devices. This data is gathered in a large volume and is termed as big data. This huge amount of data, i.e., big data, is used for various analytical purposes [5]. The two-way communication present between consumers and suppliers in SG ensures efficient utilization of resources [6]. Nowadays, the major challenges faced by SG are electricity cost minimization and user comfort maximization [7]–[9]. Therefore, efficient load forecasting is necessary for demand side management (DSM) in terms of maximizing user comfort and minimizing electricity cost [10].

Handling big data remains a crucial issue for areas like social networking, computational intelligence, machine learning, data mining, etc., [11]–[14]. Different frameworks like Apache Spark are proposed for the in-depth analysis of data. The main focus of these frameworks is the representation of data and its prediction [15]. Usually, velocity, variety, veracity and volume are considered the main characteristics of data acquired from SGs, as shown in Figure 1. For analyzing high volume of real time data, efficient processing is required [16]. Collecting data from different sources (sensors and smart meters) is considered as the first step in big data

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

TABLE 1. List of abbreviations.

Abbreviations	Description
AC	Auto correlation
AI	Artificial intelligence
AMI	Advance metering infrastructure
ANN	Artificial neural network
CBAF	Cluster based aggregation forecasting
CFS	Correlation feature selection
CNN	Conventional neural network
CRV	Cross validate filter
DBOR	Distance based outlier rejection
DFT	Discrete fourier transform
DT	Decision tree
DWT	Discrete wavelet transform
ED	Euclidean distance
EF	Ensembler filter
EMS	Energy management system
FDNN	Feed forward deep neural network
FE	Feature engineering
FS	Feature selection
GA	Genetic algorithm
GMI	Generalized mutual information
IFPM	Incremental frequent pattern mining
IS	Instance selection
KNN	K nearest neighbor
LSSVM	Least square support vector machine
MAPE	Mean absolute percentage error
MI	Mutual information
MIMO	Multiple input multiple output
MSE	Mean squared error
NB	Naive bayes
NED	Normalized euclidean distance
NN	Neural network
PAA	Piecewise aggregation approximation
PC	Principle component
PCA	Principle component analysis
PF	Partitioning filter
PSO	Particle swarm optimization
RDNN	Recurrent deep neural network
RF	Random forest
RVM	Relevance vector machine
SA	Simulated Annealing
SAX	Symbolic aggregation approximation
SD	Standard deviation
SEMS	Smart energy management system
SOHAC	Storage optimizing hierarchical agglomerative
STLF	Short term load forecasting
SVM	Support vector machine
WPT	Wavelet packet transform
WT	Wavelet transform

analytics [14]. This data may include energy consumed or demanded from SG, data sensed by sensors, smart meters' electricity consumption records, history of weather forecast, etc. However, efficient integration, storage and cleansing of data are the challenging issues. Therefore, this survey paper focuses on data preprocessing techniques, which have a great impact on the prediction model's output.

Preprocessing of data is considered as the most important task in transformation of data to an acceptable and suitable form for analysis purpose. In preprocessing, dataset size is reduced, data is normalized and outliers present in the dataset are detected. Generally, the following four steps are used in data preprocessing [17].

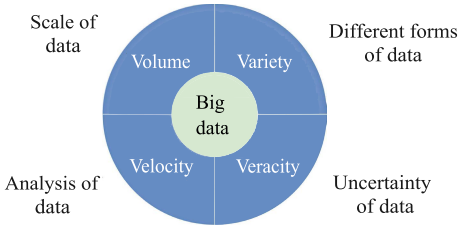


FIGURE 1. Big data characteristics.

- Data cleansing.
- Data reduction.
- Data integration.
- Data transformation.

Initially, data is analyzed in two steps: data acquisition and data integration. Firstly, data is acquired from the real world. The data may be incomplete, noisy or redundant. Therefore, data cleansing is used for removing redundancy in data to make it consistent. Missing values are filled and smoothing of data is done using clustering, binning and regression. Secondly, data taken from multiple sources such as files, databases, etc., is initially integrated into a single source. Afterwards, data is reduced and transformed in such a way that it does not alter its identity and becomes useful for further processing. Inconsistent and noisy data is removed in preprocessing, which plays a very important role in data analysis.

The road map of the paper is shown in Figure 2. The remaining paper is structured as follows. Section II presents the related work, which is further divided into two subsections: related work of survey papers and related work of technical papers. Moving ahead, Section III comprises data preprocessing steps and methods used for performing data preprocessing. Section IV presents the critical analysis of the works discussed in this survey paper. The future challenges are highlighted in Section V while the paper is concluded in Section VI. Table 1 presents the list of abbreviations. While Table 2 presents the symbols used in the equations given in the paper.

II. RELATED WORK

This section presents the related work of the existing research done in data preprocessing domain. The discussion is divided into two-subsections. One dealing with the existing survey papers and the other highlighting the work done in technical papers.

A. RELATED WORK OF EXISTING SURVEY PAPERS

In [1], a broad literature of deep learning models for power forecasting of solar panels, wind turbines and electric load forecasting is presented. In the survey, the datasets utilized for testing and training of different forecasting techniques, which allow researchers to identify suitable datasets for their studies, are discussed. The comparison of numerous deep learning models is also performed in the work. In the

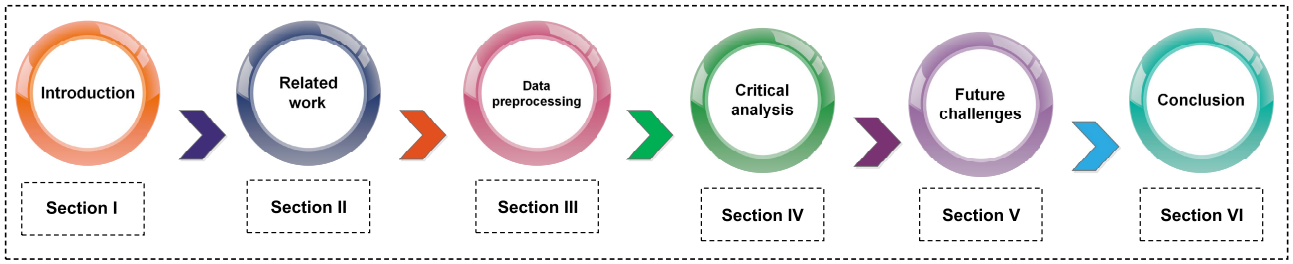


FIGURE 2. Road map of survey.

study, it is revealed that the performance of the forecasting techniques relies on the amount of available data where a huge data storage system and fast computing technologies are used to deal with the big data issues. A short review of forecasting techniques and optimization mechanisms for tuning hyperparameters is presented in [2]. Moreover, data preprocessing models are discussed. The forecasting models are compared based on their error methods, preprocessing methods and hyperparameters' optimization. Furthermore, the data preprocessing models and the existing optimization methods are critically analyzed and their important findings are highlighted. In the paper, the authors present a review of the previous survey studies and their recency scores are calculated according to the number of recently reviewed studies in them. In conclusion, the authors discuss the future research directions in details. In [3], the authors present an overview and a comparison of load management in a SG along with their technologies and development problems. Utility and consumer concerns in the context of load management are discussed to improve the intuition of readers on the topic. In the research, dynamic pricing and incentive models, which are the major categories of load management, are compared and discussed. Moreover, description of dynamic pricing schemes based on home energy management and related optimization methods, and their comparison are included in the work. It is concluded that finding an appropriate control and communication infrastructure, optimizing the consumption of energy and creating load management policies are the ongoing research domains that are related to efficient load management in the SG.

In [4], a comprehensive survey of wireless communication technologies is presented to implement a SG in a better manner. Many attributes of the network including data rate, power usage, Internet protocol support, etc., are considered for comparing the technologies in the context of the SG. The mechanisms that are appropriate for home area networks such as Z-wave, IPv6 over low-power wireless personal area networks, Wi-Fi and ZigBee are compared and discussed in the contexts of network attributes and consumer concerns. The discussion of similar methods in utility concerns' context, used for the wireless communication mechanisms such as GSM and WiMAX based cellular standards is presented. The challenges of SG applications and related network problems are discussed at the end. The authors in [5] deeply discuss many big data techniques as well as big data analytics in the SG context. Moreover, the authors present opportunities and challenges that are brought about by the emergence of machine learning techniques and big data origination from the SGs. The authors conclude that Apache Spark is more appropriate for both real time and batch processing in SGs. However, Apache Spark is a centralized storage system and therefore, it needs a third party to manage the storage system. Moreover, the authors advise users to install Apache Spark on Hadoop where the advantages given by the Spark are fully utilized together with a parallel distributed storage system.

In [6], a comprehensive survey of advancements made in machine learning models is provided. In the paper, a short survey of SG designs and their sources of data are discussed. Furthermore, data security requirements and types of false attacks are presented. In the survey, recent machine learning detection methods are summarized, which are grouped into three main detection scenarios. The groups are load forecasting, state estimation and non-technical losses. Furthermore, the authors investigate the future research directions at the end of the work by focusing on the deficiencies of the current machine learning methods. In [7], a review for deep learning models to forecast electricity load is presented. This review elaborates the previous works performed on the bases of distributed deep learning methods and conventional deep learning methods. It is concluded in the paper that data aggregation dependencies are advantageous to reduce electricity load forecasting's computational time. In [8], detailed analyses of machine learning methods used and big data originated in the energy sector are presented. Big data analysis for smart energy controls, impact, measurement, applications, operations and problems are discussed in this survey. Machine learning and big data methods are required to be used after carefully analyzing the energy system issues. Determining the match between the advantages of machine learning and big data for resolving the issues present in energy systems holds paramount importance. These methods help to operate and plan the SG or conventional power grid. In the study, the basics of machine learning and big data methods are also discussed along with their applications in various domains such as customer service, e-commerce, finance, retailers, web and digital media, telecommunication, health care and

TABLE 2. List of symbols.

Symbols	Description
Equations 1 - 4	
v_A	Mean of total inputs of the same batch
z_j	Output value
y_j	Input Value
$\bar{y}_j$	Normalized value
n	Mini batch size
s_A^2	Variance in mini batch
β, ω	Learning parameters
ϵ	Model's performance error
AN	Batch normalization function
Equations 5 - 7	
$\bar{z}$	Predicted data
z_j	Actual value
m	Total number of predicted data
DC^2	Determination coefficient
Equation 8	
$\bar{y}_j$	Normalized value
y_j	Actual value
y_{min}	Minimum value
y_{max}	Maximum value
Equation 9	
j, k	Indexes
$\bar{y}_{j'}$	Mean value
$\bar{y}_j$	Normalized value
l_j	Domain break-point
Equation 10	
$y_{1,j}$	Load data for sample 1
$y_{2,j}$	Load data for sample 2
m	Dimension
y_{max}	Maximum load of every dimension
Equation 11	
sig	Original signal
$\varphi(2^{k-1}t - m)$	Scaling function
$\Psi(2^{k-1}t - m)$	Wavelet function
Equation 12	
c_j	Correlation strength
t	Time
Y_t	Time series values
$\bar{Y}$	Mean Value
Equation 13	
Ψ_j	Digamma function
n_{neigh}	Number of nearest neighbor
M	Total number of points
$m_{p(i)}$	Number of points having distance from p_i
$m_{q(i)}$	Number of points having distance from q_i
Equation 14	
r	Number of features
f	Features
$\bar{c}_f$	Average correlation
$\bar{c}_{ff}$	Average variable to variable pairwise correlation
Equation 15	
m_{relief}	Randomly selected points
$diff$	Difference between two instances
w_f	Weight of feature
T	Nearest hit
S	Nearest miss
I	Instance
Equation 16	
$\bar{y}_j'$	Data points
C	Covariance matrix
L	Total number of data points
T	Transpose
Equation 17	
τ	Maximum dimension
U	Set of eigen vectors
Λ	Eigen value

TABLE 2. *(Continued.)* List of symbols.

Equations 18 - 20	
w_T	Weight of an attribute
U	Objective function
V	Centers of clusters
Q	Sample number
d_{jk}^2	Weighted ED
μ_{jk}	Membership degree of j with k.
N_h	Hidden layer nodes
Equation 21	
y_j	Sample
V_k	Cluster center
d_{jk}	Weighted ED
Equation 22	
P_{SR}^W	Resolution coefficient
$W(.)$	Wavelet function
R	Level
P_t	Value of price
S	Position
T	Length of series
Equations 23 & 24	
A_S	Approximate set
D_S	Detailed set
$\Psi_{SR(t)}$	Mother wavelet function
$\varphi_{SR(t)}$	Father wavelet function
p_{SR}^{ψ}	Obtained coefficient of father
p_{SR}^{φ}	Obtained coefficient of mother
S^*	Optimal position of data signal
Equations 25 - 27	
P_t	Time Series
α	Scaling parameter
k	Index
t	Time
Equations 28 - 31	
v_j^d	Distance $a_j \in \mathbb{B}$
u_j^d	Distance of $b_j \in \mathbb{B}$
$\mathbb{A}$	Represents the set of data points
$\mathbb{B}$	Uniform random distribution
$\bar{\alpha_0}, \bar{\alpha_1}$	Parameters estimators
Equations 32 & 33	
P_F	Normalized value
$\bar{F}$	Mean of attribute
F	Attribute
α_F	Standard deviation
Mx_F	Maximum attribute
Mn_F	Minimum attribute
min_F	Maximum value of an attribute
max_F	Minimum value of an attribute
Equations 34 & 35	
PC_j	Power consumption
j	Interval
no_{inter}	Number of intervals
PD	Power demand
Equations 36 & 37	
ζ	Coefficient of correlation
$y(k)$	Transformed signal
$x(n)$	Time domain signal
N	Input signal length
Equations 38 & 39	
$\lambda_i^*(k), \lambda_o^*(k)$	Sequences
$\Lambda_0^*(k), \Lambda_i^*(k)$	Gray coefficients
Δ	Difference
Γ	Correlation among data
Equations 40 - 42	
L_h	Historical actual load
GP_h	Weight of natural gas price index
GP_t^f	Actual power generation
L_t^{24}	Load observation at time $t - 24hrs$
w_1	Weight

electrical power. The opportunities and problems of machine learning and big data are presented in this work.

In [9], a broad analysis and reviews of the recently proposed works for secure data analytics in the SG platform are presented. However, achieving a secured data analytic solution for intelligent grid platforms is a critical issue. The development endeavors and existing research are not completely explored using the secured data analytic solutions in intelligent power grid platforms. Moreover, the distinctive behavior of the secured data analytic and its complications on the SG are discussed. A complete taxonomy abstraction for a novel process model that highlights many research problems like data communication, data security and privacy issues, load management and analysis, load prediction, secure load data processing and storage, and secure data collection and preprocessing, is presented. In conclusion, a case study is shown to illustrate the process model. The survey in [18] explores the prospects of analyzing and verifying information from various large storage systems and analyzing many socioeconomic aspects related to a criminal incident. Moreover, the survey categorizes patterns, analyzes outliers and designs better schemes for predicting crimes utilizing machine learning and data mining methods. In this survey [19], four feature selection (FS) models are utilized for finding important features of human resource datasets to enhance the accuracy of data classification on the employee attribution of companies. In the paper, machine learning models such as neural network, Naive Bayes (NB), gradient boosting tree, K-nearest neighbor (KNN), etc., are utilized for assessing the FS models' performance. It is concluded that, using FS models, the accuracy of data classification for the human resource datasets can be improved. In [20], a broad survey of the crow search algorithm along with its latest types, which are grouped into hybridized and modified schemes, is presented. In the paper, various uses of the crow search algorithm in many areas like distributed generation, economic dispatch, scheduling, image processing and FS are discussed. Moreover, suggestions regarding interested research domains related to crow search algorithm's hybridization, enhancement and possible new applications are made in the paper.

In [21], a contemporary survey for recent breakthroughs on utilizing, redistributing, planning and trading of energy that is harvested in the upcoming non-wired communication network, inter-operating with the intelligent power grid, is presented. This survey discusses the classical models for technologies that are related to renewable energy harvesting. The authors discuss the optimization and constrained operation of various energy harvesting platforms like multi-cell, multi-hop, multipoint-to-multipoint, multipoint-to-point and point-to-point systems. They also review information transfer and wireless energy techniques that ensure unique implementation of energy harvesting wireless communication. Finally, it is shown in the work that effective redistributions and mutual energy trading can highly minimize the energy consumption bill for providers of wireless services and decrease the energy consumption, respectively. Also, a comprehensive list of research directions is given, which needs further investigation. The authors in [22] present a brief introduction to big data services' framework and scientific computing models that include data storage and data collection. Moreover, big data analysis and processing based on various service requirements are discussed, which provide prominent data for servicing consumers. In addition, they introduce a cloud service based big data platform that gives improved performance results for large-scale data processing, analysis and storage. In conclusion, some big data systems' applications in various domains are summarized.

In [23], a comprehensive survey of 200 recently published articles is done to review the currently proposed works and other practices for machine learning models. Moreover, the trends in broad-spectrum applications of SG areas are discussed. In the paper, the rapid expansion and increasing interest in the machine learning models' applications for addressing the scientific problems of SGs from many perspectives are demonstrated. Moreover, it is shown that problems like the analysis for intelligent decision-making and high-performance data processing remain open and worthy for further research efforts. In addition, the future views for utilizing advanced communication and computing technologies such as 5G wireless networks, ubiquitous Internet of things and edge computing in the intelligent grids are elaborated. The authors concluded that machine learning is among the drivers of future intelligent energy platforms and the work gives an introductory basis for further development and exploration of associated insights and knowledge.

In [24], the authors present a survey of the intelligent power grid along with its relevant features as well as its various views on industrial energy distributions. They also discuss how the SGs' technologies have changed over time and still have prospects for evolving and strengthening the distribution platforms further. In [25], the authors provide a detailed survey of the current machine learning methods for detecting a false data injection's attack against the energy platform's state estimation technique. Moreover, data driven methods, utilizing machine learning mechanisms, are used to overcome the limitations of the conventional residual data detection methods. In [26], the authors present a detailed review of opportunities and challenges for the construction of SGs. The survey provides many challenges related to SGs' construction such as distribution of grid management, energy storage, demand response, network communications and interoperability. Moreover, a review of many global, national, regional and local opportunities related to the construction of a SG is presented.

In [27], the authors provide a systematic analysis for techniques utilized in the literature review to predict solar irradiance. The work's target is to see how the input data of meteorology, time slots, optimization, sample size and preprocessing methodology affect the complexity of models and their accuracy. Various findings and important parameters are presented in the paper. The survey gives important results based on the studied literature for selecting an

optimal model at a specific site. Also, the metrics utilized for measuring the forecasting models' efficiency are discussed. In [28], a structured review of some existing artificial intelligence (AI) models for security issues, fault detection, power grid stability assessment and load forecasting are presented. The authors also provide research problems when AI techniques are used to realize true SG platforms. Opportunities for using AI techniques to tackle SG issues are also presented. In conclusion, the AI applications can improve and enhance the resilience and reliability of intelligent power grid platforms. The researchers in [29] present a complete review for detecting and classifying power quality disturbances by elaborating AI tools and signal processing methods with their cons and pros. Moreover, automatic recognition methods are critically analyzed for different modes of operations, AI methods, FS methods, preprocessing tools and the energy input signal types (noisy/real/synthetic). Furthermore, the study gives prominent recommendations to those researchers that are interested in the domain of power quality analysis and wish to explore suitable methods for future enhancements. The summarized related work of the existing survey papers is given in Table 3.

B. RELATED WORK OF EXISTING TECHNICAL PAPERS

For efficient utilization of energy, an efficient energy management system (EMS) and a robust load forecasting model are required. In the existing literature, AI is used for handling complex problems. Different AI methods are used in [30] for short term load forecasting (STLF). In the study, the accuracy of a model is dependent on input parameters and data preprocessing methods. In retrospect, several FS methods are used for dimensionality reduction. However, FS techniques face challenges in the case of unbalanced data. An FS method based on genetic programming is proposed in [31]. The most discriminative features are selected that give optimal solutions. In addition, scores of all metrics are analyzed using which noisy, inconsistent and less important features are filtered out. Hence, processing and memory utilization are improved.

Effectiveness of features' classification can be improved by selecting relevant features from a dataset. It is necessary to remove noisy and redundant information without affecting the useful information. Hence, the main goal is to remove the maximum number of irrelevant features from the input datasets. The efficiency of a model and overfitting chances are reduced. For improving classification accuracy, a new wrapper method is proposed by [32] based on genetic programming.

In the field of pattern recognition, the extraction of meaningful information from a large amount of datasets is considered a challenging task. In a conventional way, batch processing is used for extracting features from whole dataset. However, this is not a feasible and convenient process. Hence, different preprocessing methods are proposed by researchers for feature extraction. The main focus of [33] is subspace feature extraction methods based on the loss function. A gray wolf optimizer (GWO) method is used for converting features into binary form. Binary GWO (BGWO) is proposed for the selection of features from the dataset. Different initialization methods are used for FS. Results are compared with existing schemes such as genetic algorithm (GA) and particle swarm optimization (PSO). The proposed technique outperforms the existing methods. Furthermore, for performance validation, testing dataset is used.

For knowledge extraction, an actual sample of data is obtained. Original dataset is divided into smaller datasets in data preprocessing. With the help of preprocessing, efficiency of load forecasting is improved. Sampling and FS methods, which are data reduction techniques, are applied for the preprocessing of data [34]. In another work, big data techniques are used to model the recency effect [35]. Modeling of recency effect is done on hourly basis. In the work, naive model is implemented in four different ways. However, the naive model does not show optimal results at the level of aggregation. Moreover, a regression model is used while considering daily data instead of hourly data. It shows higher accuracy in comparison to the naive model.

It is difficult to forecast the load of a single building as compared to aggregated load [36]. Several methods are used for load forecasting like support vector machine (SVM), artificial neural network (ANN), etc. However, load forecasting is not performed in an efficient way. Therefore, multi layer architecture and deep learning methods are used for single building load forecasting while considering short, long and medium term load forecasting. Conventional neural network (CNN) is used for forecasting. Results are compared with existing methods and CNN shows better performance.

Prediction of load is done using many approaches like neural network (NN), regression, etc., on a large scale. Prediction of electricity consumption in a market is done in [37]. In this work, the authors focus on residential areas for electricity load prediction. There are a large number of homes and electricity demand of every house varies from each other. Therefore, for feature relevancy, an FS approach is used, which finds correlation among features. Moreover, a cluster based aggregation forecasting (CBAF) scheme is used for load forecasting in the model.

STLF is considered a challenging job [38]. Error occurs in STLF when normal power distribution exceeds the committed power. In such cases, requirement of power is fulfilled by purchasing it from grid, which causes a significant increase in cost. Traditional methods of load forecasting always show less accuracy due to their poor performance. In comparison, AI methods are adopted, which are more suitable. Recurrent deep NN (RDNN) and feed forward deep NN (FDNN) are used for short term prediction of electricity.

The model proposed by authors in [39] is compared with the existing models in terms of accuracy and error estimation. Redundancy issues are found in data, which cause difficulty in price forecasting. To meet the electricity demand, efficient and accurate electricity demand forecasting is important. Traditional classifiers such as ANN and decision tree (DT)

TABLE 3. Summarized related work of existing survey papers.

Survey	Domain	Duration	Critical analysis	Future challenges	Significance
A comprehensive survey of deep learning based models for power forecasting of solar panels, wind turbines and electric load forecasting [1]	SG	31 years	-	✓	Benefits of deep learning models in power forecasting based on solar panels, wind turbines and load forecasting
A brief review of forecasting models and optimization mechanisms for tuning hyperparameters and data preprocessing [2]	SG	14 years	✓	✓	Benefits of efficiently tuning hyperparameters and data preprocessing
An overview and comparison of load management in a SG along with their technologies and development problems [3]	SG	24 years	-	✓	Benefits of using dynamic pricing and incentive models for load management
A comprehensive survey of wireless communication technologies to implement a SG in a better manner [4]	SG	11 years	-	✓	Benefits of implement wireless communication technology in intelligent power grids
A review on big data techniques as well as big data analytic solutions in the intelligent grids context [5]	SG	15 years	-	✓	Machine learning and big data application in intelligent grids
A broad survey of machine learning models' advances [6]	SG	15 years	-	✓	Benefits of utilizing machine learning methods in intelligent power grids
A review for deep learning models to forecast electricity load [7]	SG	16 years	-	✓	Application of deep learning methods in forecasting of electricity load
A detailed analysis for machine learning and big data energy sector [8]	SG	14 years	✓	✓	Machine learning and big data applications in a SG
A broad analysis and survey of the recently proposed works for secured data analytic in the SG platform [9]	SG	16 years	✓	✓	Importance of secure data analytic in intelligent power grids
The survey explores the possibilities of analyzing and verifying data from huge repositories and analyzing many socioeconomic factors related to crime incident [18]	Crime management	21 years	-	-	Application of big data analysis in crime management
A comprehensive overview of the crow search algorithm along with its latest variants, which are grouped into hybridized and modified versions [20]	Not specified	30 years	-	✓	Application of crow search algorithm
A contemporary survey for recent breakthroughs on utilizing, redistributing, planning and trading of energy that is harvested in the upcoming non-wired communication network interoperating with the intelligent grid [21]	SG	25 years	-	✓	The energy harvesting benefits in intelligent power grids
A brief introduction on general big data services framework and technical processing models that covered data storage and data collection [22]	Cloud computing based big data	15 years	-	✓	Big data and data analysis application in cloud based computing
A comprehensive survey of 200 recently published articles is done to review the currently proposed works and other practices for machine learning models and the trends in a broad-spectrum application of intelligent grid areas [23]	SG	11 years	✓	✓	Machine learning applications in the intelligent power grid
An overview of intelligent power grid along with its features as well as its various views on industrial energy distribution [24]	SG	11 years	-	✓	Advantages and challenges of industrial energy distribution in SGs
A detailed survey of the current machine learning models to detect a false information injection's attack against power platform state estimation technique [25]	SG	31 years	-	✓	Application of machine learning to detect false attacks in a SG
A detailed review of opportunities and challenges for construction of SGs [26]	SG	9 years	-	✓	Opportunities and challenges in constructing SGs
A systematic analysis for techniques utilized in the literature review to predict the solar irradiance [27]	SG	19 years	-	✓	Knowing the effects of input parameters of meteorology, optimization, time horizons, sample size and preprocessing methodology in terms of the complexity of models and its accuracy
A complete review for detecting and classifying power quality disturbances by elaborating AI tools and signal processing methods with their pros and cons [29]	SG	25 years	-	✓	AI tools and signal processing applications in detecting and classifying power quality

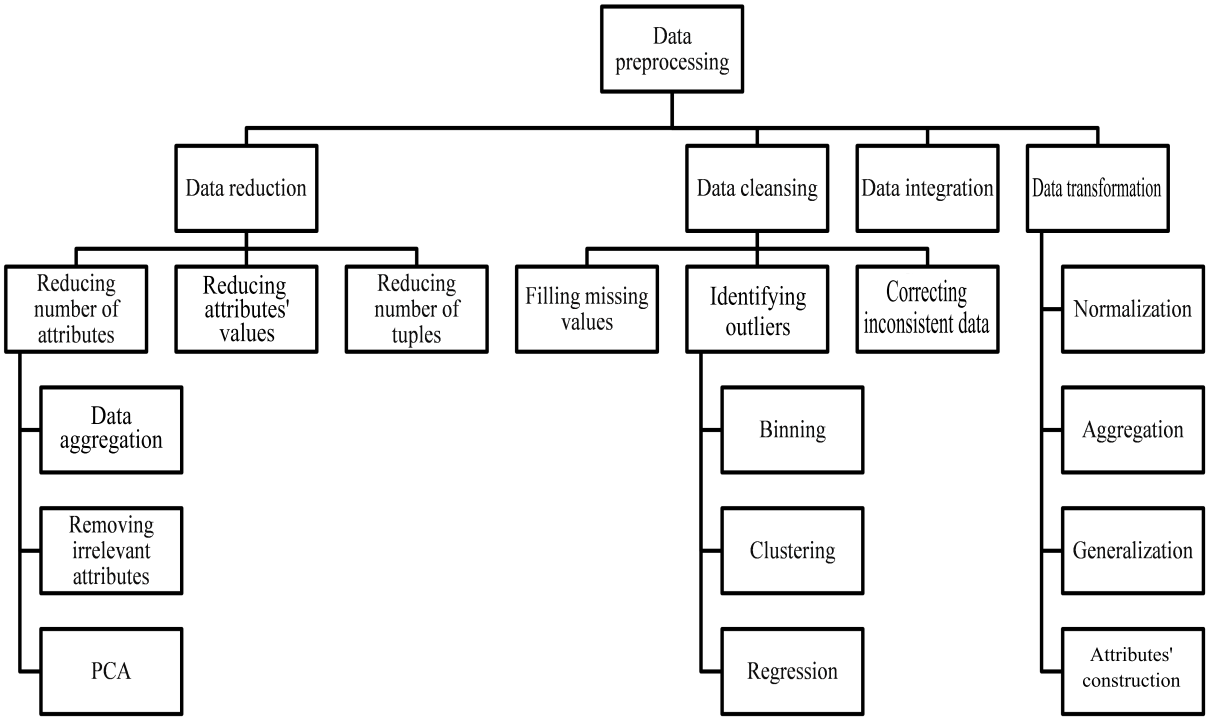


FIGURE 3. Categorization of data preprocessing techniques.

are used; however, these classifiers have overfitting problem. In the proposed model, cross validation methods are used for adjusting the hyperparameters of SVM. As the convergence rate of these methods is low, so parallel forecasting of electricity load is proposed for solving the problem.

Various layers are found in NN. The changes in parameters of previous layers in NN affect the output layers. So, there is a need of frequent changes for poor training handling. For solving such problems, batch normalization is used. The details of batch normalization are as follows [40]:

$$v_A = \frac{1}{n}\sum_{j=1}^{n} y_j, \tag{1}$$

$$s_A^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - v_A)^2, \tag{2}$$

$$\bar{y_j} = \frac{y_j - v_A}{\sqrt{s_A^2 + \epsilon}}, \tag{3}$$

$$z_j = \omega\bar{y}_j + \beta = AN_{\omega,\beta}(y_j). \tag{4}$$

z_j is output value and y_j is input value. Mini batch size is referred as n, which shows the number of inputs in every mini batch. The mean of total input for the same batch is given by v_A. Variance in mini batch is shown using s_A^2. y_j values are normalized as $\bar{y}_j$. Learning parameters are represented as β and ω. ϵ and AN are the model's performance error and batch normalization function, respectively. By applying batch normalization, efficiency of training is improved.

Nowadays, efficient price and load forecasting is a challenge due to complex and irregular nature of electricity consumption. Efficient STLF is done considering hourly price [41]. For improving accuracy of forecasting, NN is considered with multiple layers. Data is normalized before passing to the input layer. The variance is taken as 1 while mean is taken as 0. Firstly, initial weight of NN is selected considering range from $-1/\sqrt{e}$ to $1/\sqrt{e}$ where the number of neuron's input is shown using e.

$$MSE = \frac{1}{m}\sum_{j=1}^{m}(z_j - \bar{z})^2, \tag{5}$$

$$MAPE = \frac{1}{m}\sum_{j=1}^{m} \mid \frac{z_j - \bar{z}}{z_j} \mid *100, \tag{6}$$

$$DC^2 = \sum_{j=1}^{m} \frac{(\bar{z} - \tilde{z})}{(z_j - \bar{z})}. \tag{7}$$

$\bar{z}$ is the predicted value and z_j is the actual value. The total number of the predicted data is denoted as m and determination coefficient is represented by DC^2. Hence, data normalization improves the model's performance as compared to the existing approaches.

For dimensionality reduction, various methods are used in [42] like discrete wavelet transform (DWT), symbolic aggregation approximation (SAX) and discrete fourier transform (DFT). Arbitrary values ($y = y_1, y_2, \ldots, y_n$) of

consumption data are transformed into range of [0, 1].

$$\bar{y_j} = \frac{y_j - y_{min}}{y_{max} - y_{min}}. \tag{8}$$

At time t, $\bar{y_j}$, y_j are normalized and actual values, respectively. Minimum and maximum consumptions are represented by y_{min} and y_{max}, respectively. For improving potential effect of demand response, filter method is applied. Euclidean distance (ED) is calculated in time series data, which has low bounding. For reducing data dimensionality, SAX approach is used. The numerical time series data is converted into symbolic strings with the help of SAX method. Firstly, load data is transformed into piecewise aggregation approximation (PAA). Secondly, a discrete string of PAA is symbolized. In PAA, the values of amplitude that belong to identical intervals are removed. Mean values are calculated as:

$$\bar{y_j}' = \frac{1}{l_j - l_{j-1}} \sum_{k=l_{j-1}+1}^{l_j} \bar{y}_j. \tag{9}$$

The index of the normalized load data is represented by k, the index of transformed PAA load is shown with the help of j. The domain break-point at j^{th} time is l_j. At j^{th} segment, average value is represented by $\bar{y_j}'$. After applying the averaging method, smooth values are achieved. Moreover, PAA is a pruning process based on DWT, whose cost of computation is low. Discrete symbols are obtained by applying SAX algorithm.

In the classification step, a generalized normalized euclidean distance (NED) is used to find the similarity between two real-valued vectors. NED is calculated between samples of data as follows [43].

$$NED_{12} = \sqrt{\sum_{j=1}^{m} \left(\frac{y_{1j} - y_{2j}}{y_{max}}\right)^2}. \tag{10}$$

From NED_{12}, 1 and 2 represent sample 1 and sample 2, and NED is calculated between these samples. Where y_{1j} and y_{2j} represent load data for sample 1 and sample 2, respectively. Dimension is represented with m and for representing maximum load of every dimension, y_{max} is used. Hence, the application of NED ensures better understandability of data for further processing.

In power planning, analysis and forecasting of load are very important [44]. Data integration has a significant impact on prediction. Planners record the load data and perform the analysis considering various scales of time. Load data consists of various features. By considering these features, a coordinated forecasting method is proposed. In this way, accuracy and efficiency of the proposed method are improved.

Data quality is improved after applying preprocessing methods, which help in analysis of the predicted data. In [45], a hybrid approach is used for load forecasting. This hybrid approach is applied to the historical data, which is taken from real world. The proposed methodology has two steps. Firstly, for prediction of price at an individual level, relevance vector machine (RVM) is used. Secondly, aggregation is applied to individual prediction. After that, regression is applied. Results are compared with individual RVM, NB and auto regressive moving average (ARMA). The proposed methodolgy gives better prediction results as compared to other techniques.

Forecasting of electricity load is done using the proposed FS technique for next day load prediction [46]. Data dimensionality is reduced using principle component analysis (PCA). Dimensionality of time series data is reduced and only the distinct variables are used. The performance parameters of PCA are tuned using GA. After that, prediction accuracy is measured. The proposed model gives better accuracy in terms of load and price forecasting as compared to existing models. However, interdependencies between load and price data are not considered. So, an approach, termed as multiple input and multiple output (MIMO), is proposed. Correlation is calculated among electricity price and load data. Three components are considered in the proposed model. With the help of wavelet packet transform (WPT), subsets of data are made. Moreover, DWT is used as a filter method in many scientific papers. On the other hand, load signals and electricity price data have non linear patterns. Therefore, coefficient vector approximation is found in DWT. However, some useful information is lost. Hence, WPT is used for decomposition and finding approximate coefficients. The computational time of the proposed model is decreased using WPT. For selecting the best input candidate, generalized mutual information (GMI) is used. Forecasting is done using least square support vector machine (LSSVM). For simultaneous prediction of load and price, MIMO model is used as a base model for the proposed framework, LSSVM-MIMO [47]. In DWT, original signal is obtained using the following equation.

$$sig(t) = \sum_{k=1}^{m} \gamma_{k-1}(m)\varphi(2^{k-1}t - m) \; e \sum_{k=1}^{m} \lambda_{k-1}(m)\Psi(2^{k-1}t - m), \tag{11}$$

where the scaling function is given by $\varphi(2^{k-1}t - m)$ while wavelet function is given by $\Psi(2^{k-1}t - m)$. The intermittent nature of renewable energy sources makes the management of electricity prices difficult [48]. Therefore, there occurs an imbalance between electricity production and consumption, due to which the grid becomes unsteady. For removing uncertainty and ensuring grid stability, a combination of four deep learning models is used in the proposed work for predicting electricity load and improving prediction accuracy.

In this survey, the main focus is on preprocessing techniques used in big data. Data is taken from SG using different sensors and devices. Comparisons of various preprocessing

methods are made. At the end of this paper, a critical analysis is performed and future directions are provided.

III. DATA PREPROCESSING

Data preprocessing involves transforming raw data to well-formed datasets so that data analytic can be applied. Using the data preprocessing methods, irrelevant and redundant data is removed from the dataset. Data preprocessing is considered as the foremost step in data analytics. Data cleansing, integration, transformation and reduction are the most important steps of preprocessing. The main purposes of data preprocessing are to remove all irrelevant data and ensure consistency in the data representations. In Figures 3 and 4, data preprocessing hierarchy and steps involved in preprocessing big data originating from the SG are shown, respectively.

A. DATA REDUCTION

It is the transformation of digital numerical or alphabetical information derived empirically or experimentally into a corrected, ordered and simplified form [49]. Data reduction is shown with the help of Figure 5. Infrequent and large spikes are found in electricity data due to dynamic nature of its price. For electricity price prediction, better results can be achieved by ignoring spikes during the process of estimation [50]. In this way, time series data is obtained, which is easy to handle and use for electricity forecasting. Similarly, log transformation is used as a preprocessing step and for parameter estimation. To further enhance model performance, the data is normalized within [0, 1] interval using machine learning models. In this way, accurate modeling is done. However, statistical significance is not considered [51]–[53].

A filtering approach is applied on electricity consumption data to filter the data for further processing. A two step method is proposed for the selection of relevant features [54]. In the first step, daily and weekly patterns of electricity load data are captured. After that, using FS approaches, subsets of these features are formed. Four FS approaches are used to extract relevant features: mutual information (MI), auto correlation (AC), correlation FS (CFS) and RReliefF. Furthermore, the value of AC function is computed and on the basis of this function, correlation strength is identified between features using the following equation [54]:

$$c_j = c(Y_t, Y_{t-j}) = \frac{\sum_{t=j+1}^{m}(Y_t - \bar{Y})(Y_{t-j} - \bar{Y})}{\sum_{t=1}^{m}(Y_t - \bar{Y})^2}. \quad (12)$$

Time series value at time t is given as Y_t. $\bar{Y}$ represents the mean value of all values of Y in a given time slot. j is an autocorrelation coefficient and c_j is a linear correlation. Dependencies between two features Y and Z are measured using MI. In case of dependency, MI has a positive value and in case of independency, it has a negative value. MI is a very important FS method used for electricity load forecasting. Non-linear and linear correlation among features can be captured using this method. MI, based on KNN, is applied on data distribution for computing MI between two features [54].

$$MI(Y,Z) = \Psi(n_{neigh}) - \frac{1}{n_{neigh}} - \frac{1}{M}\sum_{i=1}^{M}[\Psi(m_p(i)) + \Psi(m_q(i))] + \Psi(M). \quad (13)$$

$\Psi(j)$ is digamma function, n_{neigh} is the number of nearest neighbors, and P and Q are considered as features. $m_p(i)$ and $m_q(i)$ are the number of points p_k with a distance to p_i satisfying $||p_i - p_k|| \leq \eta_p(i)/2$ and the number of points q_k with a distance to q_i satisfying $||q_i - q_k|| \leq \eta_q(i)/2$, respectively. The distance between p_i and its n_{neigh}^{th} neighbor is given by $\eta_p(i)/2$ and the distance between q_i and its n_{neigh}^{th} neighbor is given by $\eta_q(i)/2$. An individual feature subset is produced explicitly in CFS. All features are ranked individually and final subsets of variables are formed. Subsets of features are not correlated with each other, rather, they are correlated with the predicted values.

$$Merit_s = \frac{r\bar{c}_f}{\sqrt{r + r(r-1)\bar{c}_{ff}}}. \quad (14)$$

The number of features is represented by r, average correlation between every feature f is denoted as $\bar{c}_f$. c_{ff} represents average variable to variable pairwise correlation.

RReliefF is applicable for forecasting, selection and classification tasks [55]. Probability values are used in RReliefF to differentiate the values of two classes. RReliefF randomly selects an instance I and searches for its closest neighbor in the same class (nearest hit) and in another class (nearest miss). The nearest hit is represented as T and the nearest miss is represented as S. Through the following equation, weights w_f of features are updated. Difference between two instances is calculated using *diff* function [55]. The values are then normalized between 0 and 1.

$$w_f = w_f - \left(\frac{diff_f(I,T)}{m_{relief}} - \frac{diff_f(I,S)}{m_{relief}}\right). \quad (15)$$

Randomly selected variables m_{relief} from training data cause increased variations in FS method. For data reduction, all m_{relief} examples are replaced by training examples. The reliability of a feature's weight is also increased. RReliefF works well on irrelevant, noisy and redundant data. Its complexity time is linear, so, it is an efficient algorithm as compared to others. In order to perform perfect load estimation, sufficient information regarding electricity load and features of data must be available.

A dataset having millions of records has high chances of including missing and erroneous values, along with outliers. For analysis purpose, data must be in a proper format. An outlier is an unwanted training item that has an unexpected feature value due to non-ordinary conditions or exceptions [56]. The rejection of outliers is an important task. For this purpose, an outlier rejection algorithm is proposed by [57]. Distance based outlier rejection (DBOR) method is used for removing outliers. In case, outliers are present in data, the classifier

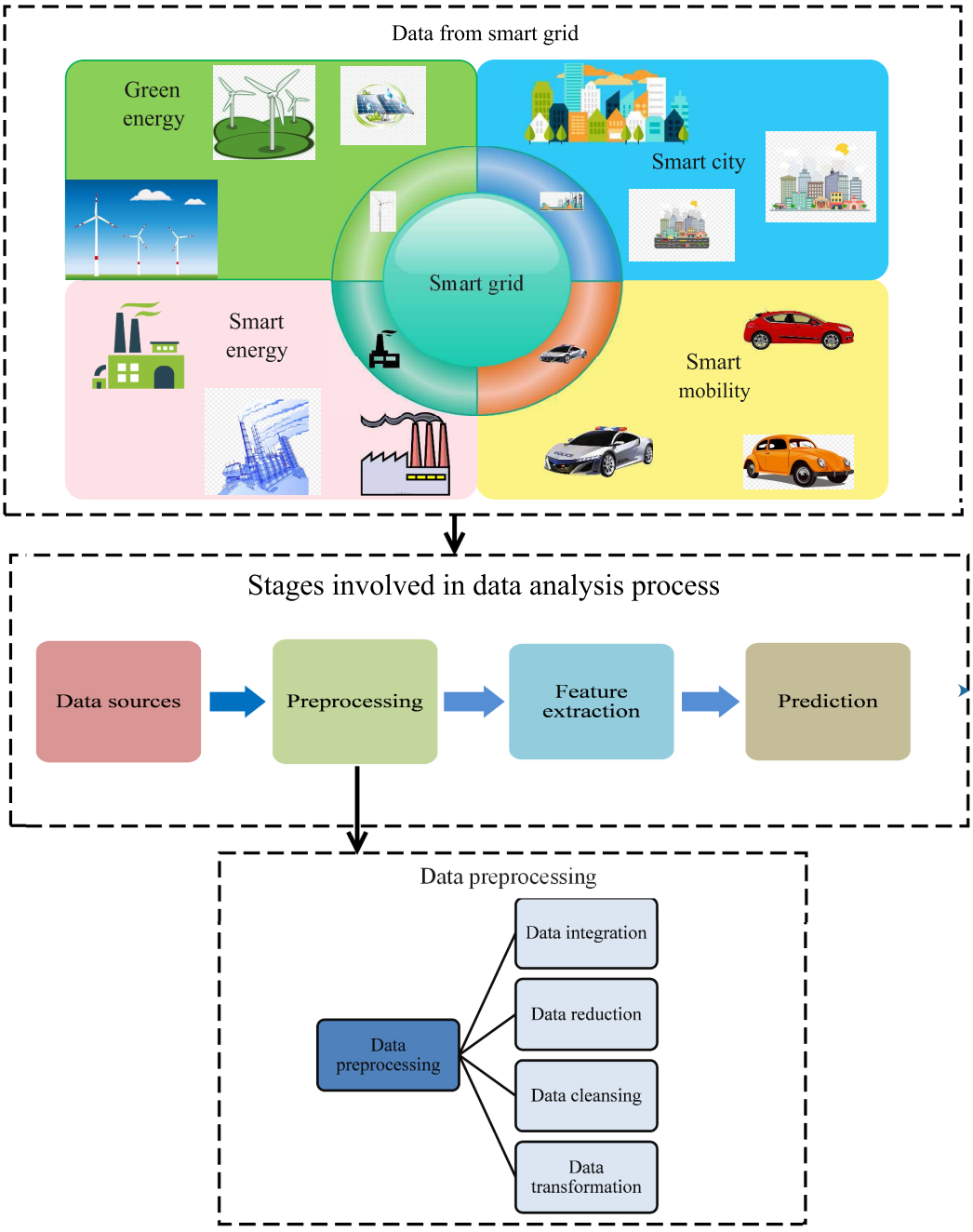


FIGURE 4. Steps involved in preprocessing big data originating from the smart grid.

may overfit. For further processing, data preparation is required. In [48], a clustering and incremental frequent pattern mining (IFPM) mechanism is proposed. In addition, the correlation of data sample is calculated between appliances' instances. Moreover, classes are constructed on the basis of members' similarity and dissimilarity in cluster analysis [48]. Using the above mentioned processes, the volume of data is reduced; however, the analytical results are the same. Data can also be reduced by the following ways.

1) NUMBER OF ATTRIBUTES ARE REDUCED

Different FS methods are used to reduce the number of attributes and select only the most prominent and useful features.

2) NUMBER OF ATTRIBUTE VALUES ARE REDUCED

Attribute values can be reduced using PCA [58]. It is used for the representation of data vectors. Firstly, normalization

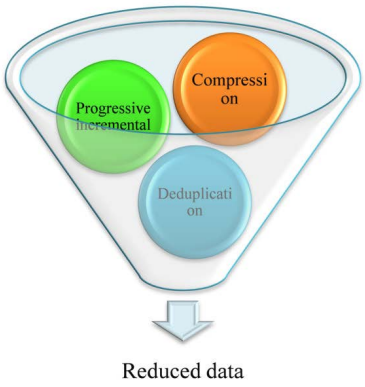


FIGURE 5. Data reduction.

is done so that all attribute values lie in the same range and no attribute has any dominance over other attributes. Secondly, vectors are calculated on the basis of normalized data. Computed vectors are known as principal components (PCs). Thirdly, these PCs are arranged in a decreasing order. In this way, patterns or groups are identified. Lastly, size of data is reduced by removing components that have very low variance and original data approximation is done.

3) NUMBER OF TUPLES ARE REDUCED

Dimensionality of data is reduced for minimizing computational complexity [59]. PCA is used for feature reduction. In this method, accuracy of clustering is improved using the following equation:

$$C = \frac{1}{L}\sum_{j=1}^{L} \acute{y}_j \acute{y}_j{}^T, \tag{16}$$

$\acute{y}_j$ are data points where $j = 1, 2, \ldots, L$ and C is a covariance matrix. The eigenvectors of C are computed as:

$$CU = U\Lambda \Rightarrow C = U\Lambda U^T = \sum_{\tau=1}^{\tau_i} \lambda_\tau \mu_\tau \mu_\tau{}^T, \tag{17}$$

$U = [\mu_1, \mu_2, \ldots, \mu_{\tau_i}]$ is a set of eigen vectors, $\Lambda = diag\ (\lambda_1, \lambda_2, \ldots, \lambda_\tau)$ is eigenvalue, τ_i is considered as maximum dimension. In PCA dimensionality reduction, weights are also assigned.

$$w_\tau = \frac{\lambda_\tau}{\sum_{\tau=1}^{\tau_i} \lambda_\tau}. \tag{18}$$

w_τ represents weight of an attribute. Objective function can be minimized by determining U and centers of clusters V.

$$\mathbb{J}(U, V) = \sum_{j=1}^{N_h}\sum_{k=1}^{Q} Q(\mu_{jk})^m d_{jk}^2, \tag{19}$$

$$0 \le \mu_{jk} \le 1, \sum_{j=1}^{N_h} \mu_{jk} = 1, \tag{20}$$

where N_h represents hidden layer nodes, Q represents the sample number and μ_{jk} denotes the membership degree of j with k. The weighted ED from sample y_j to cluster center V_k, denoted as d_{jk}, is defined as follows:

$$d_{jk} = \sqrt{\sum_{\tau=1}^{\tau_i'} w_\tau (y_{j\tau} - V_{k\tau})^2}. \tag{21}$$

Furthermore, time series data taken on daily basis should be preprocessed and sorted. So that missing, redundant and noisy values are removed. In [60], a model comprising moving average method and WT is proposed. The moving average method is used for data smoothing. In addition, preprocessing of data is done using WT while information division is done on frequency basis. In the model, price series data of electricity is divided into sub series and its coefficients are adjusted using the following equation [60]:

$$P_{SR}^W = 2^{-\frac{S}{2}} \sum_{t=0}^{T-1} P_t W\left(\frac{t - R.2^S}{2^S}\right) = 2^{-\frac{S}{2}} \sum_{t=0}^{T-1} P_t W_{SR}(t), \tag{22}$$

wavelet function is represented as $W(.)$, value of price at time t is P_t. Length of series is taken as T. Resolution coefficient is shown as P_{SR}^W, where R and S are level and position, respectively. Mother and father wavelets are based on multi-resolution methods. Low frequency is extracted using father wavelet function while high frequency is extracted using mother wavelet function. Hence, $A_S(S = 1, 2, \ldots, S^*)$ is an approximate set and $D_S(S = 1, 2, \ldots, S^*)$ is a detailed set, which are defined by the following equations:

$$A_S = \sum_R p_{SR}^{\phi} \varphi_{SR}(t), S = 1, 2, \ldots, S^*, \tag{23}$$

$$D_S = \sum_R p_{SR}^{\Psi} \psi_{SR}(t), S = 1, 2, \ldots, S^*, \tag{24}$$

$\psi_{SR}(t)$ is mother wavelet function and $\varphi_{SR}(t)$ is father wavelet function. p_{SR}^{ϕ} and p_{SR}^{Ψ} are obtained coefficients of mother and father wavelet functions, respectively and S^* is the optimal position of the data signal.

$$\varphi(t) = \sum_{k=-\infty}^{\infty} \alpha_k \sqrt{2_\alpha}(2t - k), \tag{25}$$

$$\psi(t) = \sum_{k=-\infty}^{\infty} (-1)^k \alpha_{k+1} \sqrt{2_\alpha}(2t - k). \tag{26}$$

The basic time series $P_t(t = 1, 2, \ldots, T)$ can be written as:

$$P_t = D_1 + \ldots + D_{S^*} + A_{S^*}. \tag{27}$$

B. DATA CLEANSING

Data cleansing is the process of detecting and correcting corrupt or inaccurate records [61]. A record may consist of a set, table or a database, which is used to identify incorrect, inaccurate or irrelevant parts of the data. After that, the corrupt data gets replaced, detected or modified with accurate records. The main purpose of data cleansing is the

transformation of raw data into useful information where the contents are made readable and easy to access. Data plays a very vital role in various organizations. Hence, maintenance of data is very important. Figure 6 shows the steps involved in data cleansing after it is being imported and before it is being used. Data cleansing is required before the data is used because the imported data is mostly in the form of raw data, which contains noise, outliers, redundant data, etc. Data cleansing can be done in the following ways.

1) FILLING MISSING VALUES

Data collected from real world is inconsistent due to the presence of noise and missing values. The missing values in the dataset occur due to many reasons like limited storage, disagreement in uploading the data, compromising the input devices and sometimes because of security reasons. The missing values adversely affect the reliability and performance of the machine or deep learning models. Keeping this in view, these values need to be handled before proceeding with the development of the model. Certain approaches can be adopted to fill these missing values such as calculating means of attributes, probability or sometimes by ignoring data rows. Filling the missing values would make the data consistent and noise free.

2) IDENTIFYING OUTLIERS

A sample is randomly taken from population and distance is calculated among two values. If the distance is abnormal, sample point is referred as an outlier. Otherwise, it is not an outlier. The data sample point that is far from the mean position is also referred to as an outlier.

a: METHODS TO FILL MISSING VALUES AND OUTLIERS

The following methods are used to fill missing values and outliers.

- Clustering: task of grouping objects in such a way that objects in the same group (called a cluster) are more similar to each other than those in other groups [62]. Clustering lies in the category of unsupervised classification. It is used for distribution and preprocessing of data. Large volume of data is divided into multiple groups on the basis of feature similarity [63]. Owing to the increase in population, the data associated with each electricity user is also increasing day by day, which brings a severe challenge of data storage and processing. For this purpose, storage optimizing hierarchical agglomerative clustering (SOHAC) is proposed for fulfilling storage requirement [64]. For space optimization, inconsistent and redundant records are removed.
  A MapReduce framework is used for implementation of the proposed parallel clustering approach [65]. The design of the proposed technique is based on K means clustering approach. K means clustering does not work well on a large number of datasets as compared to the approach proposed in [66]. Cluster tendency can be measured by calculating the degrees of clusters using the following equation [67]:

$$H = \frac{\sum_{j=1}^{n}(v_j^d)}{\sum_{j=1}^{n}(v_j^d) + \sum_{j=1}^{n}(u_j^d)}, \tag{28}$$

  considering that the nearest neighbor v_j^d is the distance of $a_j \epsilon \mathbb{A}$ from its nearest neighbor in A and u_j^d is the distance of $b_j \epsilon \mathbb{B}$ from its nearest neighbor in A. Here, $\mathbb{A}$ represents the set of data points in a d dimensional space and $\mathbb{B}$ represents the set of uniformly randomly distributed data points.
- Regression: It is a set of statistical processes in which relationships among variables are estimated. Predictions of a range of numerical or continuous values, which are found in specific datasets, are also taken. It can also be used as a linear regression in which relationship is estimated among two variables. There exists only one independent variable y_j to model m data points where α_0 and α_1 are considered as parameters. So, equation of straight line can be written as follows [68]:

$$x_j = \alpha_0 + \alpha_1 y_j + \epsilon_j, \quad j = 1, \ldots, m. \tag{29}$$

  Equation of multiple linear regression can be written as:

$$x_j = \alpha_0 + \alpha_1 y_j + \alpha_2 y_j^2 + \epsilon_j, \quad j = 1, \ldots, m. \tag{30}$$

  Sample linear regression model can be written as:

$$\bar{x}_j = \bar{\alpha}_0 + \bar{\alpha}_1 y_j, \tag{31}$$

  where $\bar{\alpha}_0$ and $\bar{\alpha}_1$ represent the parameter estimators.
- Imputation: two common approaches are used to fill the missing values: deleting the missing values and imputing the missing values [69], [70]. The former approach is not suitable and is never recommended because the entire row or column is deleted if the missing value is found in the dataset. In this way, important information might be lost as well. Therefore, the latter approach is commonly used to handle the missing values. The approach includes replacing missing values with arbitrary numbers, replacing with mean, median and mode, replacing with the previous and the next values using forward and backward fills, and replacing with univariate and multivariate approaches. In univariate approaches, only one feature is considered. For univariate operations, simple imputer and linear interpolation are mostly used in which the missing values are filled by taking the mean of their next and previous values. Whereas, in multivariate approaches, more than one features are taken into account. The k-nearest imputer (KNI) is one example of multivariate approaches. In KNI, initially, those records are searched that have the missing values. Then, the missing values are filled using the appropriate searched value. Moreover, the missing values in the dataset are stored as not a number (NAN). Therefore, if the data is restricted due to some security reasons, the values in the dataset are stored as NAN, which are easily tracked and

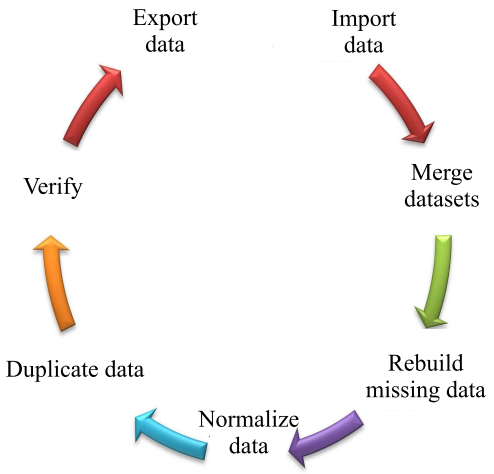


FIGURE 6. Data cleansing.

handled by the aforementioned imputation methods. The imputation methods are selected according to the nature of the data and most of the time through performing experiments. Hence, it is concluded that the discussed imputation methods perform well even when the data is restricted because of security concerns [69], [70].

3) REMOVING INCONSISTENT DATA

Data taken from the real world is in huge volumes and it is mostly inconsistent, meaning it consists of anomalies, missing values, outliers, etc. This data gives poor performance in terms of prediction and classification. Hence, the inconsistencies present in the data are removed before using it for various purposes.

4) REMOVING NOISE

The removal of noise from the data cleanses it. The following types of methods are used for removing noise.

- Filter method
- Wrapper method
- Embedded method

The three methods belong to the class of conventional FS methods. These methods are discussed as follows.

- Filter method: Subsets of features in datasets are primarily determined using filters [71]. These subsets are dependent on size of the data. Many learning algorithms like random forest (RF), Relief-F, etc., are applied on feature subsets to evaluate the type of data. RF combines the output of individual decision trees, which is a random subset of features, and generates the final output. The final output is obtained after filtering less optimal feature subsets. Relief-F, on the hand, calculates the feature score of each feature and ranks the features accordingly. The features are then filtered on the basis of rank. Moreover, proper subset selection on the basis of consistency criteria becomes a difficult task. Based on the nature of the problem, cross validate filter (CRV), ensembler filter (EF) and partitioning filter (PF) are used as per requirements [72]. In CRV, the features are divided into subsets and performance of each subset is tested. The features that give poor performance are filtered out and the best feature is selected. While, PF partitions the entire dataset in form of chunks and selects the partition on which the model performs the best. The steps of filter method are shown in Figure 7.
- Wrapper method: Subsets of variables are evaluated and variables' interaction is detected, as shown in Figure 8. In wrapper methods, the FS process is based on machine learning techniques that try to fit on particular datasets [71]. As a result, overfitting risk and computational time are increased in wrapper methods. Examples of the method are stepwise selection, backward elimination, forward selection, etc.
- Embedded method: Different techniques are used for assessing data. In training phase, gain ratio of every attribute is automatically adjusted, incomplete records are removed and system validation is done, as shown in Figure 9.

FIGURE 7. Filter method.

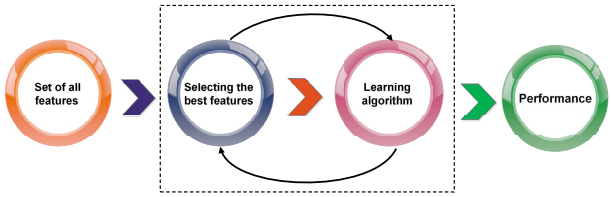


FIGURE 8. Wrapper method.

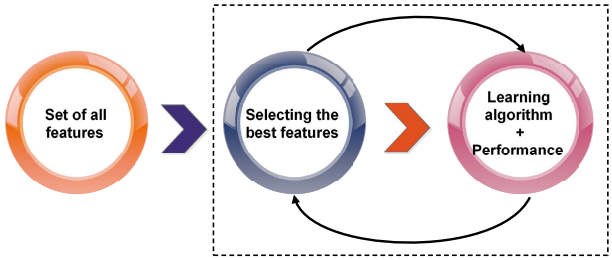


FIGURE 9. Embedded method.

The subset's size is dependent on variety and size of samples. Verification test is applied in the case of incomplete records. Hence, missing values are replaced and removed. After that, with the help of a surrogate filter, values are substituted in the primary filter. In the surrogate filter, the filtered values are taken from one filter and transferred to another filter, which is considered as the primary filter in most cases. The surrogate filter can work in the case of homogeneous data. However, it is hard to choose a surrogate filter

in heterogeneous data. It is because issue of interoperability would arise when data would be different. Hence, appropriate values are checked at the global level and a surrogate filter is chosen [74]. The above mentioned techniques are implemented on the samples of structured data, however, in case of unstructured data, these methods are not applicable.

The advantages and disadvantages of FS categories are discussed in Table 4.

C. DATA TRANSFORMATION

Data is converted from one structure to another structure according to volume, complexity and format. Transformed data can be simple or complex [73]. Different technologies and tools are used for data transformation, e.g., Talend, Pentaho, CloverDX, etc.

1) NORMALIZATION

Data is organized in a tabular form such that data redundancy and dependency are reduced [75]. Data is divided into smaller tables and relationship is defined among all tables. Data normalization can be done using the following methods: decimal scaling, z-score and min-max normalization (the most commonly used method). Attribute values are normalized by calculating standard deviation (SD) and mean in z-score normalization, given in the following equation.

$$\bar{P_F} = \frac{F - \bar{F}}{\sigma_F}, \tag{32}$$

σ_F is SD and $\bar{F}$ is mean of attribute F. When the minimum and maximum values, max_F and min_F, of attribute F are known, then z-score normalization is used. Linear transformation can be performed on data using the following equation [75].

$$P_{trans} = \frac{F_i - min_F}{max_F - min_F}(Mx_F - Mn_F) + Mn_F. \tag{33}$$

P_{trans} is calculated considering $Mx_F = 1$ and $Mn_F = 0$. Relationship among data values can be preserved using min-max normalization. Data is normalized between the range 0 to 1. In [42], [43], [76], [77], data is normalized after aggregating it on daily and hourly bases.

$$PC = \sum_{j=1}^{no_{inter}} PC_j, \tag{34}$$

PC_j represents power consumption in the j^{th} interval. The number of intervals are shown with the help of no_{inter}. Moreover, PD shows the power demand. In the proposed work, all coefficient average values can be calculated using the following equation:

$$PD = max_j(PD_j), \quad J = 1, \ldots, 24. \tag{35}$$

$$\gamma_{0_j} = \frac{1}{m}\sum_{t=1}^{m} \zeta_{0j}(t). \tag{36}$$

where $\zeta_{0j}(t)$ is the coefficient of correlation. For determining dominant frequency, DFT is used:

$$Y_{(k)} = \sum_{n=0}^{N-1} x(n)e^{-i2\pi kn/N}, \tag{37}$$

where $x(n)$ and $Y_{(k)}$ are the time domain and the transformed signals, respectively and input signal's length is represented by N. To normalize the data, equation (8) is used. After data normalization, correlation among data is given as [78]:

$$\Gamma(\Lambda_o^*(k), \Lambda_i^*(k)) = \frac{\Delta_{min} + \zeta_{dist}\Delta_{max}}{\Delta_{oi}(k) - \zeta_{dist}\Delta_{max}}, \zeta_{dist}\epsilon(0, 1), \tag{38}$$

$$\begin{aligned} \Delta_{oi}(k) &= |\lambda_o^*(k) - \lambda_i^*(k)|, \\ \Delta_{max} &= max_{i,k}|\lambda_*(k) - \lambda_i^*(k)|, \\ \Delta_{min} &= min_{i,k}|\lambda_*(k) - \lambda_i^*(k)|. \end{aligned} \tag{39}$$

$\Lambda_o^*(k)$ and $\Lambda_i^*(k)$ are gray coefficients between sequences $\lambda_o^*(k)$ and $\lambda_i^*(k)$. Distinguished coefficient is represented by ζ_{dist} and its value is set as 0.5. Data normalization is done using equation (8). Weight w is assigned to each feature n and the normalized data is represented as follows:

$$d(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + \ldots + w_n(x_{in} - x_{jn})^2}. \tag{40}$$

ED is calculated in [79] using the following formula:

$$||D|| = \sqrt{(L_h - L_t^{24})^2 + (GP_h - GP_t^f)^2}. \tag{41}$$

The historical actual load is represented by L_h. The weight of natural gas price index is represented by GP_h while GP_t^f and L_t^{24} are the actual power generation and load observed at time $t - 24hrs$, respectively. As mentioned above, data preprocessing usually includes normalization and data cleansing. Load data has some missing and defective values, which are removed and replaced by applying an averaging method. Data is also normalized between maximum and minimum values of electricity. Suppose training data has large values, then weights are adjusted according to sigmoid activation function and data is normalized [41]–[43], [80]. SAX is used for reducing datasets. The input values are transformed in the range [0, 1]. Moreover, there exists a ratio between moving average and input variable. Correlation between data is calculated on the basis of the ratio, given in [81]. This ratio is calculated using capacity utilization function (CUF):

$$CUF = \frac{RD(t)}{AC(t)} * 100, \tag{42}$$

where the residential load at time t is represented by $RD(t)$ while the available capacity at time t is represented by $AC(t)$.

Tables 5-8 present the details of different methods used in literature grouped on the basis of four different data preprocessing methods: averaging, aggregation, normalization and dimensionality reduction. The working discussed in the first column of these tables present the methods used for preprocessing data. Different works use different methods for data preprocessing. Table 9 provides the summary of different data preprocessing techniques.

TABLE 4. Advantages and disadvantages of different categories of feature selection methods.

	Types	Advantages	Disadvantages	Examples
Filter	Univariate	• Fast • Scalable • Independent of the classifier	• Ignore feature dependencies • Interaction with classifier is ignored	ED
	Multivariate	• Models feature dependencies • Independent of the classifier • Better computation complexity as compared to wrapper methods	• Slower than univariate techniques • Less scalable than univariate techniques • Ignores interaction with the classifier	CFS
Wrapper	Deterministic	• Simple • Interacts with the classifier • Models feature dependencies • Less computational complexity as compared to randomized methods	• Risk of overfitting • More prone than randomized algorithms to get stuck in a local optimum • Dependent selection of classifier	Simulated Annealing (SA)
	Randomized	• Less prone to local optima • Interacts with the classifier • Models feature dependencies	• High computational complexity • Dependent selection of classifier • Higher risk of overfitting than deterministic methods	GA
Embedded	All features	• More accurate than filter methods • Find the feature subset for the algorithm being trained • Much less prone to overfitting than both filter and wrapper	• Cannot be changed after configuration • They are hard to maintain.	RF

IV. CRITICAL ANALYSIS

In this section, based on the discussion of existing data preprocessing methods, a narrative is built and presented in the form of critical analysis. This section will help other researchers to draw future directions and also strengthen the performance of the existing preprocessing methods.

A. WPT ONLY SUITABLE FOR NON-CRITICAL DATA

WPT is a commonly used method in the literature for dividing data samples into groups. However, in this method, some necessary information is lost and the computational cost of data preprocessing is increased [79], [80]. Hence, it is suitable for problems with non-critical data. In such problems, the effect of data loss is not significant because similar data is present in the dataset. Moreover, the spike preprocessing method instead of WPT is a more suitable solution for dividing data samples into groups.

B. SELECTION OF A METHOD DEPENDS ON THE NATURE OF PROBLEM

Missing values are filled by taking an average of the existing values. However, from the literature, it is observed that for the purpose of load prediction, the improvement in forecasting accuracy is not up to the mark [82]. Instead of averaging data, pattern mining is a promising option. In this method, when a missing value is detected, the pattern of its previous and next values is mined, and the missing value is predicted based on these patterns. However, this method increases the computational overhead. Hence, the selection of an appropriate method to fill the missing values depends on the nature of the problem. If forecasting is time-critical and accuracy is the secondary priority, then the averaging method is suitable. Otherwise, if forecasting is not time-critical and accuracy is the top priority, then pattern mining is suitable to fill the missing values.

C. FITNESS CALCULATION METHOD BETTER THAN AVERAGING METHOD

To tackle the missing values in a dataset, fitness calculation of the best and the worst values has proven to be a suitable choice [54], [83]–[86]. It is observed that when the averaging method is replaced with fitness calculation values, better forecasting results are acquired. It implies that, before selecting a data preprocessing method, the nature of the data and problem should be investigated thoroughly, and then the best suitable method should be selected.

D. EFFICIENT TUNING OF HYPERPARAMETERS

In data preprocessing, clustering is done for supervised algorithms. DT is the simplest method of clustering and k-means is the commonly used clustering algorithm.

TABLE 5. Brief details of preprocessing methods in literature based on averaging.

Working/method	Achievements and features	Region	Limitations
For every hour, aggregated data is divided into demand and supply; hourly price average is applied on datasets	Daily average pricing is used for improved forecasting [35]	Europe	Potential relevant information is ignored, overfitting and complexity issues arise
For every smart meter, data is treated individually an average is calculated	For finding accuracy of a predictive model, values are calculated using the averaging method [36]	Ireland	Missing values are found in data, limited data for smart meter is utilized for forecasting
Averaging is performed for hourly and daily temperature	Missing values are found by taking average, recency effect is also calculated, which shows accuracy in forecasting [82]	USA	Model shows less accuracy
Applying equal weights to all data intervals and then taking their average	Quantile regression averaging is used for generation of probabilistic load forecasting [83]	Charlotte, USA	Probabilistic forecasting generation is not done
Training sets' samples have been taken for data analysis	Mean average is calculated on datasets for data smoothing [84]	Europe	Training time of data is increased
2,075,259 active power measurements are reduced to 34,608 by averaging method	Average of one minute resolution data is taken and data is reduced [85]	Macedonia	Only individual stations are considered for load forecasting
Maximum and minimum weights are calculated mathematically; average of maximum and minimum values are used	Experiments are performed and average values are taken, better solutions are achieved in terms of accuracy [86]	Australia	Gives only the local optimal solution; not global optimal solution
Finding linear correlation by taking average of two values	Missing points are covered by taking average of three consecutive values [54]	Australia	Only weekly forecasting is done while patterns of seasons are ignored
For every parameter, average normalized value is calculated	Mean and SD are calculated for improving fault detection [87]	USA	Anomalies on multi level are not detected in an efficient way
Data of 21 zones is considered, which consists of temperature and load data averaged on hourly basis	For each time zone temperature, simple average of every weather station is taken, which helps in improving security attacks [88]	Charlotte, USA	Less accuracy and consistency of data in load forecasting due to data integrity attacks
Averaging of three values is performed and missing values filled	Taking average of first three data values for filling missing values [97]	Ireland	Only STLF is done
Fitness is evaluated on the basis of average, best and worst values	Accuracy is improved [98]	Hebei Province, China	Paradigm of Internet energy is totally ignored
Taking average of hourly data	Taking average on datasets for accurate prediction [99]	Taiwan	Only one step ahead or two step ahead prediction is done
Scalar and hourly price is transformed into normalized data and averaging is performed	Autoregressive moving average model is used for reducing attributes and the existing systems are extended for improving results [100]	Spain and Germany	Hourly pricing is not considered

TABLE 6. Brief details of preprocessing methods in literature based on aggregation.

Working/method	Achievements and features	Region	Limitations
Four different aggregation methods are used: Previous hour method, previous day method, previous week method, previous year method	Four methods of aggregation are used [37]	ISO New England	Abnormal conditions of weather are not observed
All data is aggregated at data concentrator	Two types of data, aggregator data and disaggregator data, are used, which aggregate data in an efficient way [104]	Not mentioned	Storage and processing of data are ignored
Time of use tariff is used and aggregated hourly datasets are considered	Scale independent matrix is used for data aggregation to find consumption behavior during different time of use [105]	Ireland	Consumption of energy prediction during different time scales is not considered
Aggregation is done on consecutive three months' hourly data	Prediction error is minimized [106]	USA	Flexibility of power consumption for consumers is not considered
Datasets are divided into sub datasets on the basis of similarities found in aggregated data and DT criterion is used for data splitting	Data partitioning is done using clustering and DT [107]	Hong Kong	No suitable tools and methods for analyzing data in big buildings
Acquired data is aggregated and processed	Data is processed to design 3 experiments: Aggregated customers considering hourly data, aggregated customers considering 15 minutes data, individual household considering hourly data [109]	USA	Uncertainties of features are not captured in an efficient way
Firstly, addition is performed and secondly differential mechanism is applied	Aggregation is applied on energy consumption data [110]	National Meterological and Hydrological Institute	Privacy of data is not considered

The algorithms have hyperparameters, which need to be tuned carefully to produce accurate results. In literature, heuristic algorithms have gained popularity in finding suitable values for hyperparameters. However, the inclusion of these algorithms increases the execution time of the preprocessing methods [87]–[92].

TABLE 7. Brief details of preprocessing methods in literature based on normalization.

Working/method	Achievements and features	Region	Limitations
Data is normalized between 1 and -1	Day ahead electricity price forecasting [38]	Belgium	Training time is increased
5 customers are chosen randomly from each category of data; take mean of every mini batch and total input mean of the same mini batch	Data cleansing, normalization and structure change [39]	Korea	Overfitting of data
Data is normalized in the form of batches	Batch normalization is used [40]	North America	Time complexity is found
Data normalization is done using mean 0 and variance 1	Data is normalized [41]	Australia	Computational time is increased
Using min max formula, normalization is done and data is normalized between 0 and 1	Normalization is done [42]	Not mentioned	Filtering on base load is not done smoothly
Data is normalized after removing noise and data cleansing	Integration is performed to get normalized data and punctuality is improved [44].	Shandong, China	No balance between sub demand and aggregated demand
113 outliers are detected out of 7,604 data instances, data of remaining instances is normalized	Outlier filters are applied for values' normalization, Computational time is decreased [45]	Spain	Overfitting is found in variable selection
Maximum 15 minutes demand is taken and data is normalized	Normalization is done, 1 for event occurring and 0 for other reading [77]	Canada	Operational cost is increased
Data is normalized on daily basis by finding minimum and maximum values	Data is normalized using a formula on daily basis [42]	Not mentioned	Consumption abnormal behavior is not detected
NED is calculated between maximum and minimum values of load	NED is calculated [43]	China	Complexity is found in information
Normalization of data is performed	Price forecasting accuracy is improved [78]	England	Overfitting is found in data selection
The difference is taken between same entities and the resultant values are normalized	ED is calculated for analyzing data in an efficient way [79]	USA	Filtering of outliers is not done in an efficient way; therefore, forecasting accuracy is not improved
Input is normalized for SVR's best structure	Data is normalized [80]	Iran	Convergence criteria is changed
Minimum and maximum values are used keeping the dta in te normalized value	Data is normalized, model performed well on day ahead and week ahead forecasting [94]	England	Stability and robustness of model are not considered
Correlation is calculated between electric load, minimum, maximum, average and adjusted temperature; data is given in normalized form	Pearson correlation coefficient is used [95]	Korea	Accuracy in prediction is not improved
Datasets are analyzed in both frequency and time domain and data is normalized	Univariate DFT is used [96]	England	Dominant factors of electricity data are not highlighted

TABLE 8. Brief details of preprocessing methods in literature based on dimensionality reduction.

Working/method	Achievements and features	Region	Limitations
400 million data records are reduced to 40 million without affecting accuracy	Clustering is done on the basis of members' behavior, IFPM is used for pattern mining [48]	Canada	Accuracy of prediction is not improved for short and long term
Data is divided into blocks and redundant data is removed	MapReduce method is used for normalization, required power generation is predicted with 99 percent accuracy [51]	USA	Datasets' complexity is found due to its larger volume
Spikes in datasets are removed to reduce data dimensionality	Spike preprocessing is applied to improve the accuracy of the model [52]	Belgium	Redundant information is not considered
Filtering is applied and dimensionality is reduced; datasets are divided into data frames for further processing	Filtering is applied to remove noisy data [53]	Spain	Data handling issues occur
Outlier rejection and FS are used	Filtering of data samples is done and outliers are filtered out [57]	Egypt	Overfitting is found in FS
PCA is used for dimensionality reduction	Predictive performance is improved [59]	Australia	Complexity is found in clustering algorithm
Electricity price series is divided into sub series, which behave in an efficient way	Wavelet preprocessing is used [60]	Australia	Data becomes complex for analyzing
Correlation is found between datasets and dimensionality is reduced	Autocorrelation is used. For reducing dimension, 1 week sliding window is used [54]	Australia	Only linear dependencies are captured during prediction

E. DATA INTEGRATION MAKES DATA COMPLEX

Integration of data from several sources makes the data complex. The complexity of the data is reduced by FS, discretization or instance selection [42], [43], [77], [78]. Sometimes, this step is done manually without using any method, i.e., selecting the useful features. However, for big

TABLE 9. Summarized overview of big data preprocessing techniques.

Paper	Central tendency	SD	Quartile deviation	Dimensionality reduction	Features' transformation	Denoising	Filter	Wrapper	Embedded	MI
[36]	-	-	-	-	-	-	-	-	✓	-
[37]	-	-	-	-	-	-	-	-	✓	-
[38]	✓	✓	-	✓	✓	-	-	✓	-	✓
[39]	-	✓	-	✓	-	-	✓	-	✓	-
[42]	✓	-	✓	-	✓	-	✓	-	-	-
[44]	-	✓	-	✓	✓	-	✓	-	✓	-
[46]	✓	-	-	-	✓	-	-	✓	-	✓
[47]	✓	✓	-	✓	✓	-	-	✓	-	✓
[52]	-	✓	-	-	✓	✓	-	-	-	✓
[57]	✓	-	-	✓	-	-	-	✓	-	-
[59]	-	-	✓	-	-	✓	-	✓	-	✓
[60]	✓	-	✓	-	-	✓	-	-	✓	✓
[42]	✓	-	✓	-	-	✓	-	-	✓	✓
[79]	✓	-	✓	-	-	✓	-	-	✓	✓
[81]	-	-	-	-	-	-	-	-	✓	-
[84]	-	-	-	-	-	-	-	-	✓	-
[87]	-	-	✓	-	-	✓	-	✓	-	✓
[89]	-	✓	-	-	✓	-	✓	-	-	-
[90]	-	✓	-	✓	-	-	✓	-	✓	-
[91]	-	✓	-	✓	-	-	✓	-	✓	-
[92]	✓	-	✓	-	✓	-	✓	-	-	-
[97]	✓	-	✓	-	-	✓	-	-	✓	✓
[99]	-	✓	-	-	✓	-	✓	-	-	-
[100]	-	✓	-	-	✓	✓	-	-	-	✓
[101]	✓	✓	-	✓	✓	-	-	✓	-	✓
[102]	✓	-	✓	-	✓	-	✓	-	-	-
[103]	-	✓	-	-	✓	✓	-	-	-	✓
[106]	-	✓	-	-	✓	-	✓	-	-	-
[107]	-	✓	-	✓	-	-	✓	-	✓	-
[108]	✓	-	-	-	✓	-	-	✓	-	✓
[109]	✓	-	✓	-	✓	-	✓	-	-	-
[110]	✓	✓	-	✓	✓	-	-	✓	-	✓
[111]	✓	-	-	-	✓	-	-	✓	-	✓
[54]	✓	-	-	✓	-	-	-	✓	-	-
[112]	-	-	-	-	-	-	-	-	✓	-
[113]	✓	-	-	✓	-	-	-	✓	-	-
[114]	✓	-	-	✓	-	-	-	✓	-	-
[115]	✓	-	-	✓	-	-	-	✓	-	-
[116]	-	-	✓	-	-	✓	-	✓	-	✓

data or data with a large number of features, it is important to use a suitable method. This phase of data preprocessing is also known as data reduction. Some of the commonly used data reduction methods are discussed in this survey; however, there is still a room for improvement because size, shape and nature of data change over time.

F. NATURE OF DATASET TELLS WHAT STEPS TO USE

From the existing literature, it is obvious that all of the data preprocessing steps are not always necessary to improve the quality of the data. For example, a data of electricity load consumption consisting of four features, demand, price of electricity, fuel price and temperature, does not need a data reduction step. Similarly, every dataset does not need normalization or scaling. Hence, to choose the appropriate steps of data preprocessing for a dataset, it is important to understand the nature of the dataset and the problem to be solved.

V. FUTURE CHALLENGES

In the light of above discussed literature, the promising future research directions are presented in this section.

A. SCALING OF DATA PREPROCESSING TECHNIQUES

For accurate load forecasting, data preprocessing is an important step. From literature, it is analyzed that without preprocessing, the accuracy of a model is affected and computational cost is increased. Hence, the selection of an appropriate method for preprocessing is an important step. To reduce the data dimensionality, multiple FS methods are available for preprocessing [41]–[43], like PCA and WT. The methods are used for selecting the relevant features. However, with time, as the characteristics of data change, the existing FS models become insufficient. For example, in instance reduction, subsets of data are arranged to carry out the learning task, which does not show any significant improvement in forecasting accuracy. For obtaining subsets from a big database, it is necessary to have a complete set of instance reduction methods. Re-adjustment of these methods has been done for dealing with large-scale data. For this purpose, high computational capabilities are required. Hence, considering the importance of FS methods and the increase in the data volume, new and efficient FS methods are introduced for a better selection of the relevant features.

Moreover, data cleansing is an important phase. The dataset may have missing values, noise or irrelevant data, etc. Missing value imputation is used for replacing the missing values in a dataset. Here, to fill the missing values, the best possible values are estimated on the basis of the relationship among data. In addition, noise treatment is a complex problem in which similarity is measured among data points to identify and measure noise. Besides, a dataset may also include erroneous values, which are important to be tackled to improve the forecasting accuracy. From the literature, it is observed that error rate finding is low in medical, load and electricity forecasting, commerce, banking education, etc. Moreover, data cleansing is a challenging task in these fields. It is because a huge amount of data is generated due to the increasing population, which needs to be reduced and scaled. During scaling of big data, results' dependency, treatment of data according to data preprocessing capacity, iterative processing and parallelization possibility are the main concerns. Hence, new and efficient methods are required to tackle the aforementioned issues.

B. BIG DATA LEARNING PARADIGM

Data complexity, data security, data capture and data scaling problems are arising continuously in big data. To tackle these challenges, various data preprocessing methods have been proposed in the literature. As data is increasing day by day, therefore, storage issues are arising. Moreover, in semi-supervised learning, labeling of data is a complex task and real-time responses are required for processing large datasets. Besides, in the filter method, features are ordered on the basis of importance in a specific time; however, decision criteria for performing filtration are not decided [93]. Additionally, on a large dataset, it is hard to employ the wrapper method because this process involves a comprehensive search.

As we have discussed earlier, data is collected from multiple sources; hence, a large amount of variation is found in datasets after integration. During the preprocessing phase, an appropriate data sampling rate is assigned and cleansing of data is done on various levels. However, the processing cost increases. Furthermore, efficient learning and validation models are required. In addition, the solutions are restrictive to specific and complex predictors. Instead of providing results, solutions suddenly change direction towards used predictors. Moreover, prior knowledge is not considered in filter based methods. Hence, more time is required in finding solutions as compared to metaheuristic optimization methods. On the whole, data cleansing and filtering are of paramount importance. While considering large datasets in hybrid and embedded approaches, system complexity increases and an overfitting issue arises. These methods perform well on small datasets. However, performance is not good on large heterogeneous datasets. As data is increasing day by day, so, cleansing techniques are becoming difficult to be applied because of scalability and computational issues. Furthermore, it is difficult to apply preprocessing methods to unstructured data. Hence, new and improved methods are required for data cleansing, filtration, reduction and transformation.

VI. CONCLUSION

In this paper, a comparison is made between various data preprocessing methods. Data preprocessing is an important step for efficient load and price forecasting. The data collected from the real world is inconsistent, incomplete and noisy. Hence, data preprocessing is inevitable before forecasting to get accurate results. In literature, data preprocessing is categorized into four steps: data cleansing, data integration, data transformation and data reduction. However, it is unnecessary to implement all four steps on a dataset. The nature and type of dataset determine the data preprocessing steps. Data preprocessing aims to make the data meaningful and improve its quality. The problem of noisy data is solved in the data cleansing phase. Moreover, if data is collected from different sources, the data integration step is implemented. Additionally, data transformation and data reduction phases are used to transform data from one form to another and reduce the data size, respectively. The above discussion concludes that each data preprocessing step has its unique characteristics and significant nature. Moreover, data preprocessing steps seem to increase the computational time; however, they save a forecasting model from overfitting and underfitting issues and decrease a model's training time. Hence, from a detailed survey, it is concluded that data preprocessing is important for efficient and accurate forecasting.

REFERENCES

[1] S. Aslam, H. Herodotou, S. M. Mohsin, N. Javaid, N. Ashraf, and S. Aslam, "A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids," *Renew. Sustain. Energy Rev.*, vol. 144, Jul. 2021, Art. no. 110992.
[2] R. Khalid and N. Javaid, "A survey on hyperparameters optimization algorithms of forecasting models in smart grid," *Sustain. Cities Soc.*, vol. 61, Oct. 2020, Art. no. 102275.
[3] A. Mahmood, N. Javaid, M. A. Khan, and S. Razzaq, "An overview of load management techniques in smart grid," *Int. J. Energy Res.*, vol. 39, no. 11, pp. 1437–1450, Sep. 2015.
[4] A. Mahmood, N. Javaid, and S. Razzaq, "A review of wireless communications for smart grid," *Renew. Sustain. Energy Rev.*, vol. 41, pp. 248–260, Jan. 2015.
[5] D. Syed, A. Zainab, A. Ghrayeb, S. S. Refaat, H. Abu-Rub, and O. Bouhali, "Smart grid big data analytics: Survey of technologies, techniques, and applications," *IEEE Access*, vol. 9, pp. 59564–59585, 2021.
[6] L. Cui, Y. Qu, L. Gao, G. Xie, and S. Yu, "Detecting false data attacks using machine learning techniques in smart grid: A survey," *J. Netw. Comput. Appl.*, vol. 170, Nov. 2020, Art. no. 102808.
[7] M. Akhtaruzzaman, M. K. Hasan, S. R. Kabir, S. N. H. S. Abdullah, M. J. Sadeq, and E. Hossain, "HSIC bottleneck based distributed deep learning model for load forecasting in smart grid with a comprehensive survey," *IEEE Access*, vol. 8, pp. 222977–223008, 2020.
[8] S. R. Salkuti, "A survey of big data and machine learning," *Int. J. Elect. Comput. Eng.*, vol. 10, no. 1, pp. 575–580, 2020.
[9] A. Kumari and S. Tanwar, "Secure data analytics for smart grid systems in a sustainable smart city: Challenges, solutions, and future directions," *Sustain. Comput., Informat. Syst.*, vol. 28, Dec. 2020, Art. no. 100427.
[10] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Res.*, vol. 2, no. 3, pp. 94–101, Sep. 2015.
[11] H. Hui, Y. Ding, Q. Shi, F. Li, Y. Song, and J. Yan, "5G network-based Internet of Things for demand response in smart grid: A survey on application potential," *Appl. Energy*, vol. 257, Jan. 2020, Art. no. 113972.

[12] S. Aggarwal and N. Kumar, "Smart grid," in *Advances in Computers*, vol. 121. Amsterdam, The Netherlands: Elsevier, 2021, pp. 455–481.

[13] M. Massaoudi, H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access*, vol. 9, pp. 54558–54578, 2021.

[14] T. Ahmad, H. Zhang, and B. Yan, "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings," *Sustain. Cities Soc.*, vol. 55, Apr. 2020, Art. no. 102052.

[15] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[16] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 215–225, Apr. 2016.

[17] R. Tamilselvi, B. Sivasakthi, and R. Kavitha, "An efficient preprocessing and postprocessing techniques in data mining," *Int. J. Res. Comput. Appl. Robot.*, vol. 3, no. 4, pp. 80–85, 2015.

[18] P. Saravanan, J. Selvaprabu, A. Raj, A. A. Khan, and J. Sathick, "Survey on crime analysis and prediction using data mining and machine learning techniques," in *Advances in Smart Grid Technology*. Singapore: Springer, 2021, pp. 435–448.

[19] S. Chitra and P. Srivaramangai, "Feature selection methods for improving classification accuracy–a comparative study," *UGC Care Group I Listed J.*, vol. 10, no. 1, p. 1, 2020.

[20] Y. Meraihi, A. B. Gabis, A. Ramdane-Cherif, and D. Acheli, "A comprehensive survey of crow search algorithm and its applications," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 2669–2716, Apr. 2021.

[21] S. Hu, X. Chen, W. Ni, X. Wang, and E. Hossain, "Modeling and analysis of energy harvesting and smart grid-powered wireless communication networks: A contemporary survey," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 2, pp. 461–496, Jun. 2020.

[22] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," *J. Internet Technol.*, vol. 21, no. 2, pp. 393–405, 2020.

[23] M. S. Ibrahim, W. Dong, and Q. Yang, "Machine learning driven smart electric power systems: Current trends and new perspectives," *Appl. Energy*, vol. 272, Aug. 2020, Art. no. 115237.

[24] O. M. Butt, M. Zulqarnain, and T. M. Butt, "Recent advancement in smart grid technology: Future prospects in the electrical power network," *Ain Shams Eng. J.*, vol. 12, no. 1, pp. 687–695, Mar. 2021.

[25] A. Sayghe, Y. Hu, I. Zografopoulos, X. Liu, R. G. Dutta, Y. Jin, and C. Konstantinou, "Survey of machine learning methods for detecting false data injection attacks in power systems," *IET Smart Grid*, vol. 3, no. 5, pp. 581–595, Oct. 2020.

[26] S. R. Salkuti, "Challenges, issues and opportunities for the development of smart grid," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, p. 1179, Apr. 2020.

[27] A. Gupta, K. Gupta, and S. Saroha, "A review and evaluation of solar forecasting technologies," *Mater. Today*, vol. 47, pp. 2420–2425, Jan. 2021.

[28] O. A. Omitaomu and H. Niu, "Artificial intelligence techniques in smart grid: A survey," *Smart Cities*, vol. 4, no. 2, pp. 548–568, Apr. 2021.

[29] R. K. Beniwal, M. K. Saini, A. Nayyar, B. Qureshi, and A. Aggarwal, "A critical analysis of methodologies for detection and classification of power quality events in smart grid," *IEEE Access*, vol. 9, pp. 83507–83534, 2021.

[30] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, Oct. 2015.

[31] F. Viegas, L. Rocha, M. Gonçalves, F. Mourão, G. Sá, T. Salles, G. Andrade, and I. Sandin, "A genetic programming approach for feature selection in highly dimensional skewed data," *Neurocomputing*, vol. 273, pp. 554–569, Jan. 2018.

[32] B. Xue, M. Zhang, and W. N. Browne, "A comprehensive comparison on evolutionary feature selection approaches to classification," *Int. J. Comput. Intell. Appl.*, vol. 14, no. 2, pp. 142–146, 2015.

[33] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, Jan. 2016.

[34] R. Thakur and A. R. Mahajan, "Preprocessing and classification of data analysis in institutional system using Weka," *Int. J. Comput. Appl.*, vol. 112, no. 6, pp. 1–3, 2015.

[35] E. Raviv, K. E. Bouwman, and D. van Dijk, "Forecasting day-ahead electricity prices: Utilizing hourly prices," *Energy Econ.*, vol. 50, pp. 227–239, Jul. 2015.

[36] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015.

[37] A. A. Aydarous, M. A. Elshahed, and M. A. M. Hassan, "Short term load forecasting as a base core of smart grid integrated intelligent energy management system," in *Proc. Int. Conf. Mod. Electr. Energy Syst. (MEES)*, Kremenchuk, Ukraine, Nov. 2017, pp. 192–195.

[38] J. Lago, F. De Ridder, P. Vrancx, and B. De Schutter, "Forecasting day-ahead electricity prices in Europe: The importance of considering market integration," *Appl. Energy*, vol. 211, pp. 890–903, Feb. 2018.

[39] S. Ryu, J. Noh, and H. Kim, "Deep neural network based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, Dec. 2016.

[40] P.-H. Kuo and C.-J. Huang, "An electricity price forecasting model by hybrid structured deep neural networks," *Sustainability*, vol. 10, no. 4, p. 1280, Apr. 2018.

[41] H. Mosbah and M. El-hawary, "Hourly electricity price forecasting for the next month using multilayer neural network," *Can. J. Electr. Comput. Eng.*, vol. 39, no. 4, pp. 283–291, 2016.

[42] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016.

[43] P. Zhang, X. Wu, X. Wang, and S. Bi, "Short-term load forecasting based on big data technologies," *CSEE J. Power Energy Syst.*, vol. 1, no. 3, pp. 59–67, Sep. 2015.

[44] Y. Fu, D. Sun, Y. Wang, L. Feng, and W. Zhao, "Multi-level load forecasting system based on power grid planning platform with integrated information," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 933–938.

[45] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, "Short-term load forecasting in a non-residential building contrasting models and attributes," *Energy Buildings*, vol. 92, pp. 322–330, Apr. 2015.

[46] O. Abedinia, N. Amjady, and H. Zareipour, "A new feature selection technique for load and price forecast of electrical power systems," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 62–74, Jan. 2017.

[47] H. Shayeghi, A. Ghasemi, M. Moradzadeh, and M. Nooshyar, "Simultaneous day-ahead forecasting of electricity price and load in smart grids," *Energy Convers. Manage.*, vol. 95, pp. 371–384, May 2015.

[48] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, p. 452, Feb. 2018.

[49] H. Wang, Z. Yemeni, W. M. Ismael, A. Hawbani, and S. H. Alsamhi, "A reliable and energy efficient dual prediction data reduction approach for WSNs based on Kalman filter," *IET Commun.*, vol. 15, no. 18, pp. 2285–2299, Nov. 2021.

[50] C. López, D. Wang, Á. Naranjo, and K. J. Moore, "Box-cox-sparse-measures-based blind filtering: Understanding the difference between the maximum kurtosis deconvolution and the minimum entropy deconvolution," *Mech. Syst. Signal Process.*, vol. 165, Feb. 2022, Art. no. 108376.

[51] M. N. Rahman, A. Esmailpour, and J. Zhao, "Machine learning with big data an efficient electricity generation forecasting system," *Big Data Res.*, vol. 5, pp. 9–15, Sep. 2016.

[52] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Appl. Energy*, vol. 221, pp. 386–405, Jul. 2018.

[53] J. F. Torres, A. M. Fernández, A. Troncoso, and F. Martínez-Álvarez, "Deep learning-based approach for time series forecasting with application to electricity load," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.*, Corunna, Spain. Cham, Switzerland: Springer, Jun. 2017, pp. 203–212.

[54] I. Koprinska, M. Rana, and V. G. Agelidis, "Correlation and instance based feature selection for electricity load forecasting," *Knowl.-Based Syst.*, vol. 82, pp. 29–40, Jul. 2015.

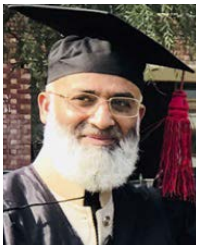
[55] A. Stief, J. R. Ottewill, and J. Baranowski, "Relief F-Based feature ranking and feature selection for monitoring induction motors," in *Proc. 23rd Int. Conf. Methods Models Autom. Robot. (MMAR)*, Aug. 2018, pp. 171–176.

[56] N. Maksaei, A. Rasekh, and B. Babadi, "Influence measures and outliers detection in linear mixed measurement error models with ridge estimation," *Commun. Statist.-Simul. Comput.*, pp. 1–17, May 2021.

[57] A. I. Saleh, A. H. Rabie, and K. M. Abo-Al-Ez, "A data mining based load forecasting strategy for smart electrical grids," *Adv. Eng. Informat.*, vol. 30, no. 3, pp. 422–448, Aug. 2016.
[58] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "Feature selection for classification using principal component analysis and information gain," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114765.
[59] Y. Lu, T. Zhang, Z. Zeng, and J. Loo, "An improved RBF neural network for short-term load forecast in smart grids," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Shenzhen, China, Dec. 2016, pp. 1–6.
[60] M. Rafiei, T. Niknam, and M.-H. Khooban, "Probabilistic forecasting of hourly electricity price by generalization of ELM for usage in improved wavelet neural network," *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 71–79, Feb. 2017.
[61] Z. Li, J. Liu, Y. Lin, and F. Wang, "Grid-constrained data cleansing method for enhanced busload forecasting," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.
[62] S. Askari, "Fuzzy C-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113856.
[63] T. Baker, "Designing and managing big data how are you researching your outcomes," in *Proc. SimTecT*, 2014.
[64] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," in *Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence)*, Sep. 2014, pp. 59–65.
[65] N. K. Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework," *J. Big Data*, vol. 2, no. 1, pp. 1–18, Dec. 2015.
[66] W. Zhao, H. Ma, and Q. He, "Parallel k-means clustering based on mapreduce," in *Proc. IEEE Int. Conf. Cloud Comput.* Beijing, China: Springer, Dec. 2009, pp. 674–679.
[67] *Cluster Analysis*. Accessed: Oct. 4, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Cluster_analysis
[68] *Regression Analysis*. Accessed: Oct. 4, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Regression_analysis
[69] W. Young, G. Weckman, and W. Holland, "A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits," *Theor. Issues Ergonom. Sci.*, vol. 12, no. 1, pp. 15–43, Jan. 2011.
[70] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," in *Proc. Int. Conf. Data Sci. Eng. (ICDSE)*, Aug. 2016, pp. 1–5.
[71] N. L. da Costa, M. D. de Lima, and R. Barbosa, "Evaluation of feature selection methods based on artificial neural network weights," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114312.
[72] S. H. Koppad and A. Kumar, "Application of big data analytics in healthcare system to predict COPD," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Nagercoil, India, Mar. 2016, pp. 1–5.
[73] A. Pereira and A. Proença, "HEP-frame: Improving the efficiency of pipelined data transformation & filtering for scientific analyses," *Comput. Phys. Commun.*, vol. 263, Jun. 2021, Art. no. 107844.
[74] R. P. Biuk-Aghai, W. T. Kou, and S. Fong, "Big data analytics for transportation: Problems and prospects for its application in China," in *Proc. IEEE Region Symp. (TENSYMP)*, May 2016, pp. 173–178.
[75] D. Singh and B. Singh, "Feature wise normalization: An effective way of normalizing data," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108307.
[76] J.-S. Chou and N.-T. Ngo, "Smart grid data analytics framework for increasing energy savings in residential buildings," *Autom. Construct.*, vol. 72, pp. 247–257, Dec. 2016.
[77] K. Grolinger, A. L'Heureux, M. A. M. Capretz, and L. Seewald, "Energy forecasting for event venues: Big data and prediction accuracy," *Energy Buildings*, vol. 112, pp. 222–233, Jan. 2016.
[78] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Y. Zomaya, "Robust big data analytics for electricity price forecasting in the smart grid," *IEEE Trans. Big Data*, vol. 5, no. 1, pp. 34–45, Mar. 2019, doi: 10.1109/TBDATA.2017.2723563.
[79] L. Wang, Z. Zhang, and J. Chen, "Short-term electricity price forecasting with stacked denoising autoencoders," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2673–2681, Jul. 2017.
[80] A. Kavousi-Fard, H. Samet, and F. Marzbani, "A new hybrid modified firefly algorithm and support vector regression model for accurate short term load forecasting," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 6047–6056, Oct. 2014.
[81] D. Keles, J. Scelle, F. Paraschiv, and W. Fichtner, "Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks," *Appl. Energy*, vol. 162, pp. 218–230, Jan. 2016.
[82] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: A big data approach," *Int. J. Forecasting*, vol. 32, no. 3, pp. 585–597, Jul. 2016.
[83] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic load forecasting via quantile regression averaging on sister forecasts," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 730–737, Mar. 2017.
[84] A. Ghasemi, H. Shayeghi, M. Moradzadeh, and M. Nooshyar, "A novel hybrid algorithm for electricity price and load forecasting in smart grids with demand-side management," *Appl. Energy*, vol. 177, pp. 40–59, Sep. 2016.
[85] K. Amarasinghe, D. L. Marino, and M. Manic, "Deep neural networks for energy load forecasting," in *Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2017, pp. 1483–1488.
[86] L. Xiao, J. Wang, R. Hou, and J. Wu, "A combined model based on data pre-analysis and weight coefficients optimization for electrical load forecasting," *Energy*, vol. 82, pp. 524–549, Mar. 2015.
[87] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5820–5830, Nov. 2018.
[88] J. Luo, T. Hong, and S.-C. Fang, "Benchmarking robustness of load forecasting models under data integrity attacks," *Int. J. Forecasting*, vol. 34, no. 1, pp. 89–104, Jan. 2018.
[89] S. Aman, M. Frincu, C. Chelmis, M. Noor, Y. Simmhan, and V. K. Prasanna, "Prediction models for dynamic demand response," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm), Data Manage., Grid Anal., Dyn. Pricing*, Miami, FL, USA, Nov. 2015, pp. 1–6.
[90] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Appl. Energy*, vol. 205, pp. 654–669, Nov. 2017.
[91] F. Javed, N. Arshad, F. Wallin, I. Vassileva, and E. Dahlquist, "Forecasting for demand response in smart grids: An analysis on use of anthropologic and structural data and short term multiple loads forecasting," *Appl. Energy*, vol. 96, pp. 150–160, Aug. 2012.
[92] P. Vrablecová, A. B. Ezzeddine, V. Rozinajová, S. Šárik, and A. K. Sangaiah, "Smart grid load forecasting using online support vector regression," *Comput. Electr. Eng.*, vol. 65, pp. 102–117, Jan. 2018.
[93] J. Chen, Z. Lyu, Y. Liu, J. Huang, G. Zhang, J. Wang, and X. Chen, "A big data analysis and application platform for civil aircraft health management," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data (BigMM)*, Taipei, Taiwan, Apr. 2016, pp. 20–22.
[94] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, p. 1168, Aug. 2017.
[95] J. Moon, K.-H. Kim, Y. Kim, and E. Hwang, "A short-term electric load forecasting scheme using 2-stage predictive analytics," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Shanghai, China, Jan. 2018, pp. 219–226.
[96] G. M. U. Din and A. K. Marnerides, "Short term power load forecasting using deep neural networks," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Santa Clara, CA, USA, Jan. 2017, pp. 594–598.
[97] Y. Li, P. Guo, and X. Li, "Short-term load forecasting based on the analysis of user electricity behavior," *Algorithms*, vol. 9, no. 4, p. 80, Nov. 2016.
[98] Z. Zhou, F. Xiong, B. Huang, C. Xu, R. Jiao, B. Liao, Z. Yin, and J. Li, "Game-theoretical energy management for energy internet with big data-based renewable power forecasting," *IEEE Access*, vol. 5, pp. 5731–5746, 2017.
[99] H.-H. Chang, W.-Y. Chiu, and T.-Y. Hsieh, "Multipoint fuzzy prediction for load forecasting in green buildings," in *Proc. 16th Int. Conf. Control, Autom. Syst. (ICCAS)*, Colorado Springs, CO, USA, Oct. 2016, pp. 1–3.
[100] J. P. Gonzalez, A. M. S. Roque, and E. A. Perez, "Forecasting functional time series with a new Hilbertian ARMAX model: Application to electricity price forecasting," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 545–556, Jan. 2018.
[101] H. Chitsaz, P. Zamani-Dehkordi, H. Zareipour, and P. P. Parikh, "Electricity price forecasting for operational scheduling of behind-the-meter storage systems," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6612–6622, Nov. 2018, doi: 10.1109/TSG.2017.2717282.
[102] M. Mao, Y. Wang, Y. Yue, and L. Chang, "Multi-time scale forecast for schedulable capacity of EVs based on big data and machine learning," in *Proc. Energy Convers. Congr. Expo. (ECCE)*, 2017, pp. 1425–1431.

[103] Y. R. V. Prasad and R. Pachamuthu, "Neural network based short term forecasting engine to optimize energy and big data storage resources of wireless sensor networks," in *Proc. IEEE 39th Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Taichung, Taiwan, Jul. 2015, pp. 511–516.
[104] R. Shyam, H. B. B. Ganesh, S. Kumar, P. Poornachandran, and K. P. Soman, "Apache spark a big data analytics platform for smart grid," *Proc. Technol.*, vol. 21, pp. 171–178, Jan. 2015.
[105] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Santa Clara, CA, USA, Oct. 2015, pp. 879–887.
[106] J. Saez-Gallego, J. M. Morales, M. Zugno, and H. Madsen, "A data-driven bidding model for a cluster of price-responsive consumers of electricity," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 5001–5011, Nov. 2016.
[107] F. Xiao, S. Wang, and C. Fan, "Mining big building operational data for building cooling load prediction and energy efficiency improvement," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Hong Kong, May 2017, pp. 1–3.
[108] A. Garulli, S. Paoletti, and A. Vicino, "Models and techniques for electric load forecasting in the presence of demand response," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1087–1097, May 2015.
[109] N. Bassamzadeh and R. Ghanem, "Multiscale stochastic prediction of electricity demand in smart grids using Bayesian networks," *Appl. Energy*, vol. 193, pp. 369–380, May 2017.
[110] V. Tudor, M. Almgren, and M. Papatriantafilou, "Employing private data in AMI applications: Short term load forecasting using differentially private aggregated data," in *Proc. Int. IEEE Conf. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People, Smart World Congr. (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, Toulouse, France, Jul. 2016, pp. 404–413.
[111] H. Chitsaz, H. Shaker, H. Zareipour, D. Wood, and N. Amjady, "Short-term electricity load forecasting of buildings in microgrids," *Energy Buildings*, vol. 99, pp. 50–60, Jul. 2015.
[112] C.-N. Yu, P. Mirowski, and T. K. Ho, "A sparse coding approach to household electricity demand forecasting in smart grids," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 738–748, Mar. 2017.
[113] Y. Chen, H. Tan, and X. Song, "Day-ahead forecasting of non-stationary electric power demand in commercial buildings: Hybrid support vector regression based," *Energy Proc.*, vol. 105, pp. 2101–2106, May 2017.
[114] N. Singh, S. R. Mohanty, and R. Dev Shukla, "Short term electricity price forecast based on environmentally adapted generalized neuron," *Energy*, vol. 125, pp. 127–139, Apr. 2017.
[115] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "Demand forecast of PV integrated bioclimatic buildings using ensemble framework," *Appl. Energy*, vol. 208, pp. 1626–1638, Dec. 2017.
[116] C. Tong, J. Li, C. Lang, F. Kong, J. Niu, and J. J. P. C. Rodrigues, "An efficient deep model for day-ahead electricity load forecasting with stacked denoising auto-encoders," *J. Parallel Distrib. Comput.*, vol. 117, pp. 267–273, Jul. 2018.

TURKI ALI ALGHAMDI received the bachelor's degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, the master's degree in distributed systems and networks from the University of Hertfordshire, Hatfield, U.K., in 2006, and the Ph.D. degree from the University of Bradford, U.K., in 2010. He is currently a Professor with the Computer Science Department, Faculty of Computer and Information Systems, Umm Al-Qura University (UQU), Makkah, and the Founding Director of the SMarT Laboratory. He has more than 15 years of research and development, academia and project management experience in IT. He holds CDCDP and CDCMP certificates. He is passionate about developing the translational and collaborative interface between industry and academia. His research interests include wireless sensor networks, energy and QoS aware routing protocols, network security, the IoT, and smart cities.



NADEEM JAVAID (Senior Member, IEEE) received the bachelor's degree in computer science from Gomal University, Dera Ismail Khan, Pakistan, in 1995, the master's degree in electronics from Quaid-i-Azam University, Islamabad, Pakistan, in 1999, and the Ph.D. degree from the University of Paris-Est, France, in 2010. He is currently a Professor and the Founding Director of the Communications Over Sensors (ComSens) Research Laboratory, Department of Computer Science, COMSATS University Islamabad, Islamabad Campus. He is also working as a Visiting Professor with the School of Computer Science, University of Technology Sydney, Australia. He has supervised 146 master's and 27 Ph.D. theses. He has authored over 900 articles in technical journals and international conferences. His research interests include energy optimization in smart/micro grids and in wireless sensor networks using data analytics and blockchain. He was a recipient of the Best University Teacher Award (BUTA'16) from the Higher Education Commission (HEC) of Pakistan, in 2016, and the Research Productivity Award (RPA'17) from the Pakistan Council for Science and Technology (PCST), in 2017. He is an Associate Editor of IEEE Access and an Editor of *Sustainable Cities and Society* journals.