Multimedia Tools and Applications (2021) 80:19675–19693
https://doi.org/10.1007/s11042-021-10696-4

Prediction of domestic power peak demand and consumption using supervised machine learning with smart meter dataset

R. Geetha [1] · K. Ramyadevi [1] · M. Balasubramanian [1]

Received: 12 July 2020 / Revised: 31 January 2021 / Accepted: 10 February 2021 /
Published online: 1 March 2021

Abstract
The prediction of electricity consumption is a vital foundation for smart energy management. Since the consumption of power varies with different appliances, better forecasting of power and peak demand is an essential accomplishment for the proper planning and development of the power generation and distribution system. This forecast analysis helps the service providers and the government to understand the lifestyle of the customers. The existing prediction and forecasting models are not meeting the standard requirements and moreover difficult to apply in practice. The forecast says that the boom of electric vehicles will increase the demand of electricity globally by 3% for the upcoming year. There exists a number of machine learning algorithms for classification and decision making. But the accuracy of the exiting methods have shown inferior performance in terms of prediction which leads to inefficient decision making in the quantity of electricity generation. This paper proposes the use of random forest supervised learning model to forecast the consumption of power and identify the level of peak demand. The large smart meter dataset collected at varying seasons of the year is fed to the random forest classifier technique for better analysis and forecasting. This approach outperforms in terms of accuracy, stability and generalization. In addition, this paper investigates the existing models and compares the performance with those models. The performance analysis shows that this model performs better than the other investigated models with performance accuracy of 95.67% and enhanced accuracy of precision and recall.

Keywords Electricity-consumption · Random forest · Artificial neural network (ANN) · Support vector machine (SVM)

✉ R. Geetha
geetha@saec.ac.in

[1] Department of Computer Science and Engineering, S.A. Engineering College, Chennai, India

1 Introduction

Electricity is well connected to human beings in day to day life, so the need for electricity is also increased. The electricity swallowed by the buildings is 30–45% of intercontinental electricity consumption. The users of electricity should know how much energy absorbed by them for every month to use the available energy effectively. According to the recent studies from World Bank, population of urbanization raises by about 3.9% and consumption of electricity raises by 1100 kWh in last ten years. Prophecy of electricity is important for consumers, energy planners. Forecasting easily helps to determine the usage of electricity by the consumers. There exist some forecast models over the consumption of electricity for buildings based on the machine learning methodologies. The machine learning has two inherited characteristics: training and interference. The excess amount of data in a training model is divided into training, testing and validation. The accuracy of training sets is validated by the testing sets. The given input data is trained using machine learning algorithm and the output obtained is called as interference. Based on the input given, training is categorized into Supervised learning, Unsupervised learning, Reinforcement learning [10, 11].

The classification part of the algorithm categorize different data sets into respective sets. Some common examples are speech recognition, hand writing recognition, face recognition, identifying spam in email etc. The types of classification algorithm are linear classifiers, support vector algorithm, k-nearest neighbor, decision trees, and random forest. Regression model creates a prophecy of target, based on independent variables. It gives a graph of turn on variable (y) and individualistic variable (x).The types of regression algorithm are linear regression, logistic regression, and polynomial regression and step wise regression n.

In unsupervised learning, the datasets are given as input but the goal or intended value is not given. It performs dimensionality reduction, clustering and density estimation. Reinforcement learning is the intermediate between the supervised and unsupervised learning. Some real-life examples of machine learning are virtual assistants like Siri, Alexa, and Google, learning through video surveillance, Social Media Services like Email Spam, Malware Filtering, suggestions and Online Customer Support. In this approach, prediction is used while computing and the existing ANN model is trained to accomplish the Ontario electricity market to provide its high potential [1, 23].

The functional regression methods are collated with the original dataset [29]. An advanced self-adaptive method named radial basis function (RBF) neural network is proposed and trained by fuzzy c-means [8]. To examine customer electricity consumption behaviors, a flexible mass-feature K-means-affinity propagation (AP) clustering algorithm is handed-down [14]. To reduce the arithmetic complexity of functional principal component analysis (FPCA), an recursive dynamic factor analysis (RDFA) algorithm is handed-down which further tracks and predicts recursively using Kalman filter (KF) [31]. An accurate temperature forecast used by a refinement model can reduce the prediction error of the electricity prices remarkably [22, 30]. To provide relative accuracy of experimental results, the most perfect is selected from other predictions. [34].The models are constructed and randomly selected commercial buildings are tested using monthly landlord utility bills. Radial basis function (RBF) with selection of kernels depends on stepwise searching method and C, γ and ε are parameters to investigate the performance of SVM. Low coefficient of variance with a low percentage error is resulted [17, 26]. The extreme machine learning and non-dominated sorting genetic algorithm use ELM based forecasting method optimization. This approach provides reliability and sharpness [3]. To solve problems of single network structure and hyper

parameter selection, a different LSTM model is used. This method divides the analyzed data based on parameter and preforms optimization [37].

A novel non-parametric approach is projected for modelling and manifold learning methodology for analysis of electricity price curves [4]. To operate smart home, real time electricity scheduling for home is needed. This method improves utilization of renewable energy. The proposed system is used to achieve the minimizing cost payment [22]. This method is proposed with the smart meter data set. It deals with utilities, usage, needs and suitability for different programs [20, 28]. The Random forest or Random decision forest is a type of algorithm that integrates few machine learning techniques into a single foretelling model. A heaps of decision trees are assembled at drilling time and outputs are in the form of classification and regression in discrete trees. Random Forest is a scheme of superintend machine learning algorithm based on combo learning. The combo learning also called as Ensemble learning that employs manifold learning algorithm to pick up a better predicted execution. Random forest is the most powerful algorithm which integrates the outcomes of various learning algorithms resulting as trees of a forest. This novel ensemble method is used for prophecy of home appliances, it integrates machine learning and statistical model for high accuracy of electricity consumption.

The motivation behind this work is that the forecast says that global electricity demand will increase by 5% for the upcoming years due to the launch of electric vehicle production hubs throughout the globe by various electric vehicle manufacturers. So there is a need for the government to forecast the electric power consumption accurately to make decisions on the quantity of electricity to be generated to outfit the customers demand.

1.1 Significance of the proposed work

Since the usage of appliances including electric car increases day by day, there is a drastic increase in the usage of electricity by the users. Hence there is a need for the accurate prediction and forecast for the sufficient generation of electric power for distribution.

The important contribution of this paper is to propose an efficient method for the accurate prediction and forecast of the monthly, weekly and daily electric power consumption using machine learning techniques to satisfy the requirements of the modern lifestyle of the customers. This forecast in turn will reduce the gap between the demand and supply thereby improving customer's pleasure. The accuracy of forecast of this method is comparatively superior to the existing methods in the literature.

The major contribution of this paper includes

- An efficient supervised learning model is applied for the efficient prediction and forecast of energy consumption to improve the satisfaction of customers with modern lifestyle.
- It addresses the use of two classes or multi class prediction problems.
- It supports mixture of categorical and continuous variables and has good prediction for data with more variables.
- It takes cares of missing data in an effective manner.
- It has the ability to handle thousands of input variables without variable selection.
- They also offer a superior method for working with missing data.
- It can automatically balance data sets, when a class is more infrequent than other classes in the data
- The accuracy of prediction of the chosen model is superior.

The rest of the paper is organized as follows: In section 2, related work carried out in the literature is presented. Section 3 describes the methods and materials proposed. In section 4 the experiments and results are presented, at the end the concluding remarks is presented.

2 Related work

This section addresses the research carried out in the forecast of the demand of the electric power by various researchers in the literature. Finally we conclude that our proposed method has superior rate of prediction than the existing proposed methods.

In this paper [23] the authors propose the idea about energy saving, load balancing. Smart grid is installed which provides the quantity of electricity. Two types of MEP models are used which analyse and determine the quantities of electricity and pricing. The suitable MEP method is used based on the situation where it checks the vulnerability and duplicate data. This model is limited to user's charge.

In recent days, the development of renewable energy for household has become a great demand. Finding the electricity usage and optimizing it, became the complicated task. In this paper [29], the authors employ statistical methods such as Gaussian distribution and Kullback-Leibler divergence. This model permits to find similarities between the patterns. Here large dataset of 500 house consumption is used. It deals with two concepts, electricity production and consumption. In this paper [8], electricity market liberation process which acts as a key driver is used. The main challenge of a Nigerian electricity sector is to provide an explanation to key. In This paper [14] the new model for Nigerian industry is used which address the current challenges but creates a new structure for Nigeria to form a secure energy future. The analysis of data is carried out in two stages. Hourly total consumption of electricity provides hourly weather related and illumination related electricity consumption. By subtracting the above two parameters, residual consumption is obtained. In second stage, agent based analytical tool is used. This tools performs many operations including optimization. A set of patterns are used to minimize negative effects on high peak demand. This work [31] deals with two major business challenges, inability to be stored economically and requirement for instant response. To compute capacity and energy charges, two mathematical equations are used. These equations can also be applied to find yearly average incentives and penalties. The use of proposed equation is associated with the unaccountable fluctuation. The primary aim is to provide benefits to the customer. In this paper [30], smart grid is used to detect abnormal electricity consumption behaviorists and cluster the similar user located in some area. This clustering is done based on electricity consumption. The electricity consumption of similar users are obtained with density clustering. The matching degrees are calculated based on historical data. Finally, abnormal electricity consumption is found with support degree. This method can effectively identify the abnormal consumption. Unsupervised learning of abnormal electricity consumption behavior is proposed [34]. The original data set is constructed by brainstorming method. Optimal feature set is selected based on variance and similarities between them. Unsupervised clustering is used to detect abnormal electricity consumption behavior. To perform evaluation label information of abnormal behavior is obtained by integrating the actual electricity consumption. This method has a benchmark on good and effective result [19]. It uses evolution based characteristics of smart meter data which removes irrelevant data and features. To predict the number of clusters, a visualization is required. K-means algorithm is used for segmentation. This method is applied on Guangdonprovice,

China. This new clustering approach provides a good segmentation of data. This paper can be used in the field of data science [5]. It deals about the process of minimizing the electricity consumption and maintenance costs. Here, the use of virtual machine and cloud data centers are necessary. Many VMs are developed and processing costs are noted. Material based fatigue model calculates the maintenance and electricity costs data center algorithm is used. This algorithm focuses on load balancing and energy consumption. It is able to accomplish peak load shifting and decrease the bill by around 12% in a typical day. In this work, [26] forecasting short range electricity price is proposed. This paper deals with an alternative method named Levenberg-MarquardtBP(LMBP) method for regular Back Propagation (BP) method. This method increases the convergence speed which is used for training ANN model by MATLAB. This provides high performance and capability in forecasting short range electricity prices. This paper [3] integrates the extreme machine learning and non-dominated sorting genetic algorithm. This system uses ELM based forecasting method optimization. This approach provides reliability and sharpness. This system provides an accuracy bet.

ween 80% to 90%. This method [36] has been verified through the data provided by Australian electricity market. It forecasts the electricity spike using data mining and gives the occurrence of price. The method of predicting the occurrence of spike has not yet been discovered. This paper uses spike value prediction technique and comprehensive tool for price spike forecasting. The market data is used to test this method. It uses forecasting strategy based high resolution data. The hourly data is collected from available market. The proposed method [6] has ability to detect price spike and several price variation. To get an accurate update, it uses an intra hourly rolling framework. Here the Ontario's electricity market data is used to evaluate performance. This method is applicable to small scale storage system. The predicted result is applied to optimization platform for operation scheduling of a battery energy system. It finds the solution by using heterogeneous structure LSTM for single network. This [37] method divides the analyzed data based on parameter and preforms optimization. Finally integrates and forecasts the output. Sequence model based optimization verifies decomposed reconstructed electricity price data This methods provides accuracy and stability. The analysis of electricity price time series provides a switching nature. It provides discrete changes in competition strategies, which represents a dynamic models of Markov chain. A hidden Markov model [12] is analyses and forecasts the electricity price. The input under different scenarios are found and characterized as more relevant. This method results with good accuracy. Conditional probability transition Matric finds the probabilities of remaining in existing state. This method has been tested in Spanish electricity market. It [4] uses novel non-parametric method for modelling and manifold learning methodology for analysis of electricity price curves. Here LLE is experimented to be an efficient way for extracting the intrinsic dimension structure of electricity price curves. This method fails for long period predictions. This method provides accuracy which is verified by data taken from Eastern US. This paper [27] forecasts the use of electricity tariff. This method provides a baseline consumption and deviation from anticipated baseline. First cost game is induced by single tariff and cost minimizing is done. The polynomial time algorithm is used to compute and validate this approach. This method can be used in large scale dataset as it provides a good result. This method improves the performance. In this paper [22], real time electricity scheduling to operate smart home. This method improves the utilization of renewable energy. The proposed system is used to achieve the minimizing cost payment. This optimization problem has been solved by genetic algorithm. The proposed approach improves the performance of home scheduling. Electricity prices varies every time. In this work [24], scheduling

problem which is naturally lob as a Markov decision process is proposed. It provides the data in numerical form. This method is tested with real price data and provide economic advantages to consumer. This method provides good result for short tasks. This method [20] is proposed with smart meter data set. It deals with utilities, usage, needs and suitability for different programs. Defining and describing different customer segments will furnish decision makers with information. It not only deals in pricing and program marketing but also in resource allocation and program development. Lifestyle of customer, establish their electricity data and separate them into groups. Finally segmentation result is carried out based on energy program.

The shambled sequence is introduced in support vector regression (SVR) algorithm and evolutionary algorithm, that not only improve the prophecy accurately but also avoids converging prematurely. The electric load is subject to changes, due to cyclic economic activities or seasonal nature. The chaotic genetic algorithm is applied to improve the prophecy and genetic algorithm is applied to avoid premature converging. Both algorithms are used to determine the parameters for SVR algorithm [16].The artificial bee balcony algorithm with seasonal recurrent support vector regression model is applied to an electric load forecasting model that improves forecasting performance and functional optimization to overcome premature local optimum [15]. This paper presents a hybrid model that combines support vector regression (SVR), Empirical Decomposition Mode (EDM), the Krill Herd and chaotic mapping functions. EMD is used to decompose the input data series and SVR is used to forecast separately. KH is used to select the parameters for SVR and chaotic Mapping is used to prevent premature converging and to improve the accuracy of the whole model [35]. The SVR model is combined with differential Empirical Decomposition Model (DEMD) and Auto Regression (AR) for electrical load forecasting. The differential EMD is used to decompose several detailed parts with high frequencies and approximate part of low frequencies. The results illustrate forecasting with accuracy and interpretability [9]. A ship motion time series (SMTS) exhibits under the effects of periodic wave and strong nonlinearity. SMTS owing to wind, ocean currents and the load of ship itself, which make accurate forecasting difficult. Due to strong non linearity, the LSSVR model is used to forecast the accuracy. The chaotic cloud particle swarm optimization(CCPSO) algorithm is introduced to optimize the parameters of the LSSVR model [21]. Quantum computing mechanism is used to quantamize dragonfly behaviour to enhance the finding performance of the dragonfly algorithm, namely QDA. It conducts the data pre-processing by the complete ensemble empirical model decomposition adaptive noise (CEEMDAN) which is useful to improve the forecasting accuracy. Thus, a new electric load forecasting model, the CEEMDAN-SVRQDA model, that combines the CEEMDAN and hybridizes the QDA with an SVR model.It is proposed to provide more accurate forecasts [32]. A novel hybrid algorithm, cuckoo search and Differential evolution (CSDE) is used to solve the constrained engineering problems. CS has powerful ability on worldwide search and less control parameters, but suffers premature convergence and lower the density of population. DE specializes in local search and good robustness but both gave satisfied results. It divides the work into two groups and algorithms, CS and DE are applied independently and these groups exchange information. It provide premature convergence, stabilize the quality of solution and the computation consumption which provide satisfactory worldwide optima [33].. Nowadays Deep learning has been used in many fields like Traffic crowd, image processing speech recognition etc. Machine Learning will handle complex data but it learns from

that data whereas Deep learning can take its own decision from the data [13].To predict accuracy in traffic flow convolutional neural network is used [2]. This prediction method can be applied to other factors like weather, social and electricity The Tabular sketch of the literature review has been shown in Table 1.

The related review reveals that there is a need for an efficient algorithm in predicting the energy consumption that will help the government to plan the generation of electric power. In this paper, Random forest learning model is proposed which is superior to the methods proposed in the earlier literature in terms of the accurate prediction of the demand of customers. Random forest is usually much faster than non-linear SVM. SVM works with specific dataset and not suited for large data. Random forest is suitable for multi class problem and has many decision trees so accuracy will be high when compared to SVM.ANN are more complex in adjusting the weights and moreover.

3 Methods and materials

The proposed framework involves data preparation to load the raw data for processing and data preprocessing to eliminate the redundant data, fill the missing values etc. These steps are commonly required by the machine learning algorithms ANN, SVM and Random Forest to analyze the customers past data usage to predict the future requirement. Since the random forest algorithm is able to build a number of decision trees and the final output is based on the majority voting,it shows pleasing results in prediction than the other two methods.

Table 1 Tabular sketch of the literature review

Reference	Methodology/Algorithm	Energy Type	Application	Region
Jie Lin et. Al. [23]	Energy saving and load balancing	Electricity.	Smart buildings	–
Hideitsu et.al. [29]	Clustering method for electricity consumption pattern analysis in household	Electricity.	Household electricity	–
Norbert et.al. [8]	K-means-affinity propagation (AP) clustering algorithm. Idea about electricity consumption	Electricity.	Smart buildings	Nigeria
Wei Zhang et.al [34]	Unsupervised detection of abnormal electricity consumption	Electricity	Smart buildings	–
Jui-Sheng et.al. [7]	Hybrid Machine Learning	Electricity	Air- Conditioners in office	USA
Sumedhasharma et.al. [28]	Smart meter data set which deals usage, needs and suitability for different programs	Energy	Smart buildings	Singapore
TesfahunMolla, et.al. [25]	Cost-effective Energy Management System	Electricity	Smart Home Appliances	USA
NoorollahKarimtbar, et.al. [18]	predicting electricity energy consumption using data mining techniques	Electricity	Buildings	Iran
Jiechen et.al. [4]	non-parametric approach is projected for modelling and manifold learning methodology is used for analysis	Electricity	Buildings	–
Ranjbar et.al. [26]	Radial basis function (RBF) is used. C, γ and ε are parameters to investigate the performance of SVM	Electricity	Buildings	–

3.1 Data preparation

If the data collected contain missing values that may lead to inconsistency. Electricity consumed data must be preprocessed to upgrade the performance of the algorithm. The attributes that are remarkable are meter id, appliances usage, bill amount and units consumed. In order to fill missing data, interpolate() function is used. Based on the tie-up among attributes, data preprocessing in data mining is most time swallowed process. At the outset, the attributes which are important to make an electricity unit and amount prediction is found by attribute-evaluator and ranker as the search-method.

The following graph in Fig. 1 shows the missing values in the television and air condition data and how it recovers its missing value.

In Fig. 1 the white shaded portion shows the missing data in the TV and AC. This is due to the place is blank space or having some duplicate value or none (nan). It is identified by missing no library package.

In Fig. 2 the missing values are filled by the interpolate function. It takes the two data from the dataset and fill it by taking the average of them.

3.2 Data preprocessing

Validation techniques used in machine learning are to get the error rate, it is closer to the true error rate of a dataset. Validation technique may not be needed if volume of data is large enough to represent the population but in real world scenario there is no true volume of data representation. While tuning model hyper parameters data sample are given to an unbiased evaluation of the model.

The datasets are loaded into the library packages for analyzing the identifier by data shape, data type and estimate the missing values and duplicate values. The proposed model can be evaluated for making the best utilization of test datasets and validation. Data cleaning / preparing by renaming the given dataset is to analyze the uni-variate, bi-variate and multi-variate process. The procedures and techniques for cleaning the data differ depending on a dataset. The primary goal of data cleaning/validation is to detect and remove errors and abnormality to improve the value of the data in analytics and decision making. The dataset collected for forecasting electricity unit and price is segregated into training set and test set.

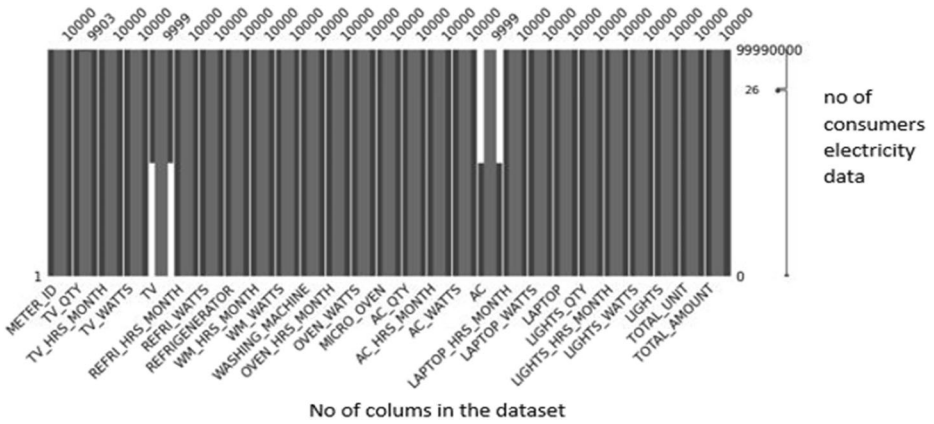

Fig 1 Identifying empty values in dataset

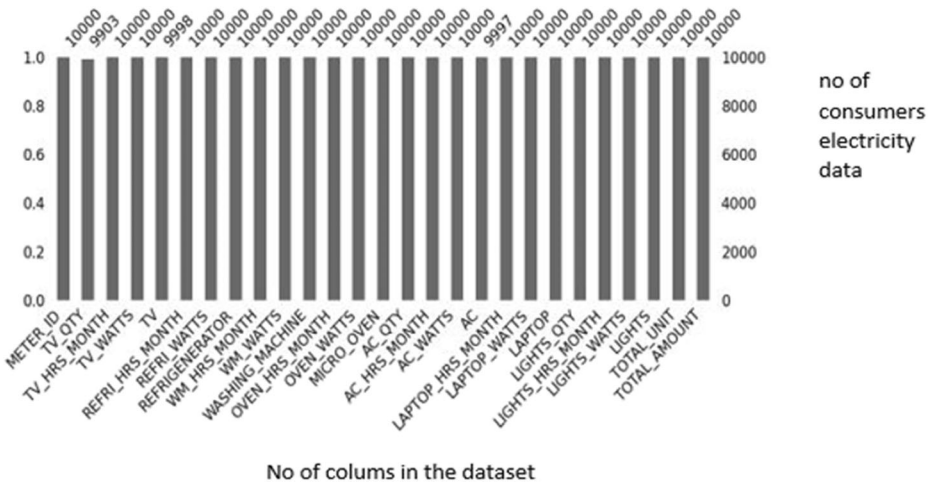

Fig. 2 Filling the missing values in dataset

Basically, to separate the training set and test set 7:3 ratio is applied. Data model is generated by using Random Forest algorithm.

3.2.1 Training the dataset

Iris data set is imported by the initial line that is predefined in module named sklearn and the table that contains information about different varieties termed as datasets.For example, to import the dataset the data_dataset variable in the load_data() function is used to enfold the program by the use of train_test_split class from sklearn package and numpy of python. Further divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values. Then the dataset is segregated as training data and test data in the 70:30 ratio. Then the algorithm is encapsulated and training data is fitted into this algorithm so that by this data the computer can be trained. At the moment, training part is complete.

3.2.2 Testing the dataset

The dimensions features helps to prophesize the species of the features using the forecast method which takes the dataset as input and separates out the forecasted target value as output.

Therefore, the output forecasted target value becomes Zero. The test score is found by the ratio of number of predictions found right and total predictions observed and accuracy score method is found by comparing the actual values of the test set with the predicted values.

3.3 Proposed method

The newly proposed method for prophecy of appliances electric consumption and weather data are integrated in classification concept of machine learning. The ultimate aim of the classification to forecast the consumption of electricity and its price and there are seven appliances data as shown in Fig. 3.

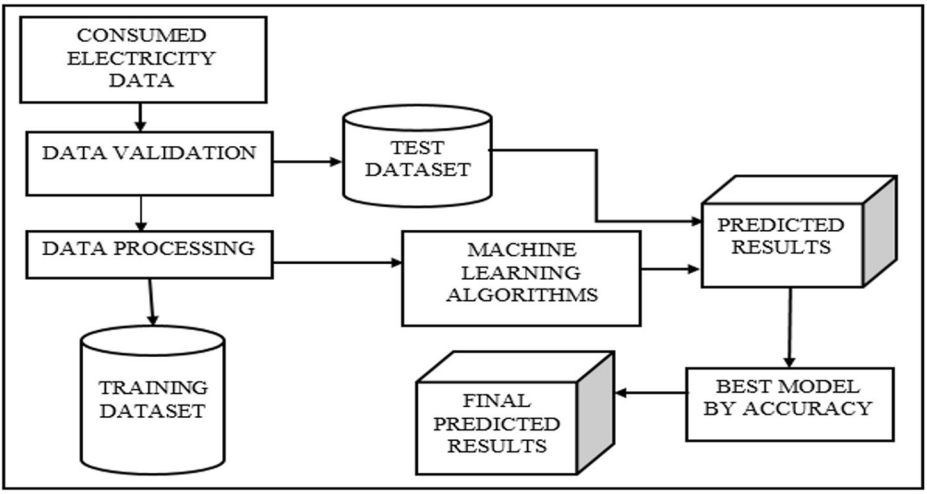

Fig. 3 Architecture of proposed method

The diagram clearly explains the users consumed electricity which is given in the dataset and then data processing is performed i,e, filing the missing data, removing the unwanted data and cleaning the data. After preprocessing machine learning algorithm (in our case random forest algorithm) is applied which trains the machine to predict and test the predicted results and validate it. If the accuracy of machine learning algorithm is less than the expected level again then the machine is again trained. Figure 4 shows the entire classification technique right from preprocessing till decision making or classifying.

3.3.1 Support vector machine

SVM belongs to the category of supervised machine learning method. It aims to work on classification of linear data initially and later on works with multidimensional data

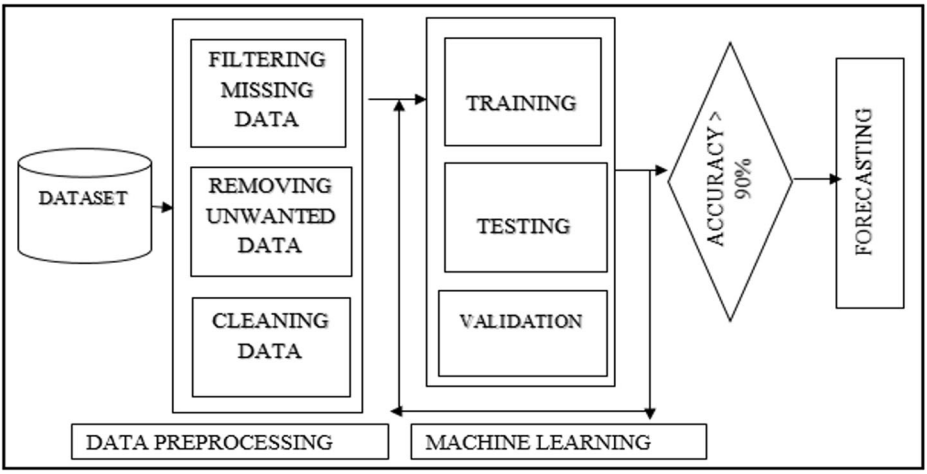

Fig. 4 Classification technique

classification. It is also capable of solving regression problems named as Support Vector Regression that is based on support vectors which is a function as shown in Eq. 1.

$$x = f(y) = U^R \phi(y) + a \tag{1}$$

where a is a constant, ϕ is any nonlinear function with the parameters x, y U and R that is used to map between the inputs and output.

Hyperplane is a slope that helps in the classification of data. Figure 5 shows the hyper plane that is illuminated as W.X + b, where X is feature vector,w is the weight vector, x is the input vector and b is the bias

3.3.2 Artificial neural network

ANN is another machine learning technique that is able to solve complex problems. It is capable of developing problems related to nonlinear category of classification and regression. Even though many variants of advanced ANN like Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) exists in the literature, we focus on the usage of basic category of feedforward neural network since the advanced techniques are more complex and require more computational power. A minimum of three layers are present in feedforward network namely input layer, hidden layer and output layer. The input layer is responsible for receiving the input data and prepares it to feed to the hidden layers. The responsibility of the hidden layer is to process the data fed by the input layer and given to the next hidden layer or the output layer. Finally the output layer combines the results received from the hidden layer to produce the final desired output as shown in Fig. 6. The output O is given as a function of inputs O as shown in Eq. 2.

$$O_i = f\left(W_{i,j} * I_j + W_{i,k} * I_k + W_{i,l} * I_l\right) \tag{2}$$

where f(x) is a threshold or activation function that stimulates the output with weight W.

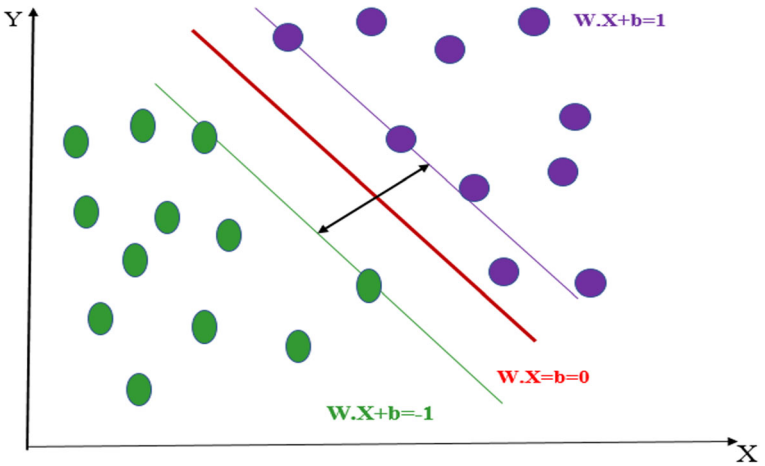

Fig. 5 Support vector machine

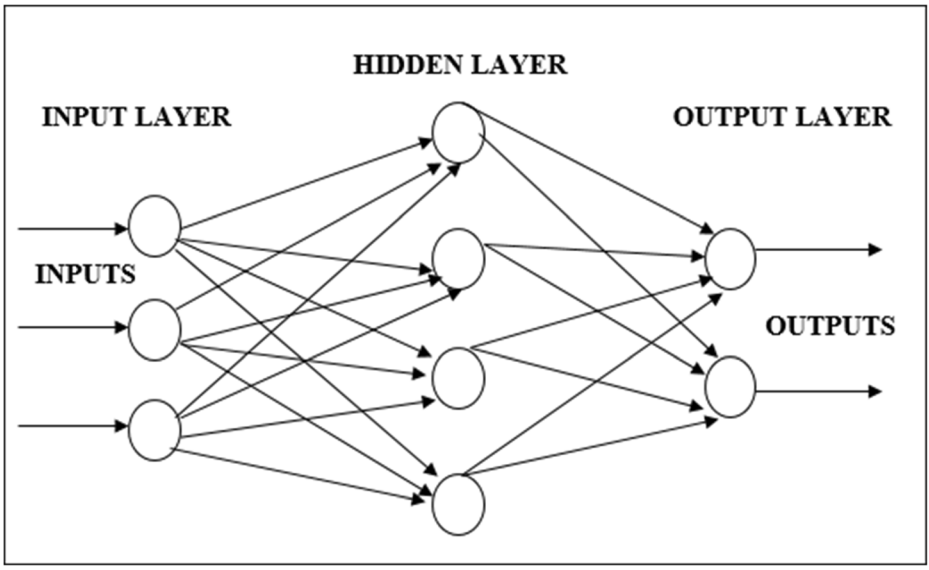

Fig. 6 Artificial neural network

3.3.3 Random Forest

The Random forest or Random decision forest is a type of an algorithm that integrates few machine learning techniques into a single foretelling model. Random Forest uses Ensemble Learning technique and it works based on the bagging algorithm. It combines the output of all the trees which is created on the subset of the data. By this, it reduces overfitting issue in decision trees and also decreases the variance and therefore increases the accuracy. A heaps of decision trees are assembled at drilling time and outputs are obtained in the form of classification and regression in discrete trees. It is the most powerful algorithm which integrates the outcome of various learning algorithm resulting as trees of a forest.

Random forest is able to build a number of decision trees and the final output is based on the majority voting as shown in Fig. 7. The regression predictor with N trees is shown as shown in Eq. 3.

$$f(y) \quad = \frac{1}{R} \sum_{n=1}^{R} \left(T_{dt} \left(y \right) \right. \tag{3}$$

$$Y = \{y_1, y_2, y_3, \ldots\ldots, y_n\}$$

where y is the n dimensional vector of inputs and $T_{dt}(y)$ refers to the decision trees.

The basic steps involved in execution of the random forest algorithm is:

Step 1: N records are chosen from the absorbed electricity data.
Step 2: Decision trees are built based on the record.
Step 3: The number of trees required by algorithm is chosen and repeat the steps 1 and 2.
Step 4: If we need to solve regression problem, the consumption is prophesized by each tree in forest.

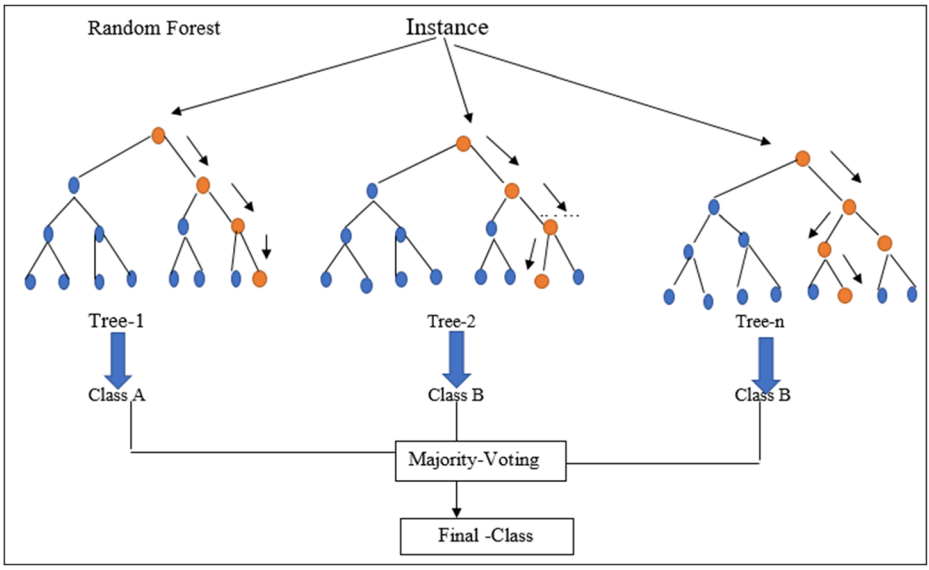

Fig. 7 Random forest

Step 5: The average of all the values produced by the decision trees is taken to calculate the final value and to prophesied.

Step 6: If we need to solve a classification problem, a new record is chosen based on majority vote and all trees in the forest prophesies the group to which the new record belongs.

4 Experiment and result

4.1 Experimental setup

The sample dataset consist of meter id and different appliances like AC, light, Cellar, Washing machine, Oven, laptops, AC and Fridge which are used by customers in day to day activities as shown in Table 2.

Figure 8 shows the mean consumption of the power. Any of the electrical appliances (like AC) run for 1 h continuously it consumes 1 unit.

Table 2 Power consumption data

Appliance	Attributes						
	Total	Daily	Weekly	Spring	Summer	Autumn	Winter
TV	0.70729	0.7.801	0.92799	1.23808	1.21797	0.89368	4.19617
Light	0.01061	0.13912	0.17024	0.31864	0.31123	0.1384	1.13491
Washing Machine	0.0898	0.00731	0.007247	0.06797	0.02696	0.02229	0.10402
Oven	0.04443	0.0039	0.3899	0.01887	0.01318	0.01177	0.05181
Laptop	0.02666	0.04831	0.02841	0.97956	0.09837	0.08968	0.48146
AC	0.17002	0.3962	0.37528	0.44499	0.4398	0.36874	1.52876
Fridge	0.82591	0.0646	0.62242	1.39698	1.40739	0.6195	4.42327

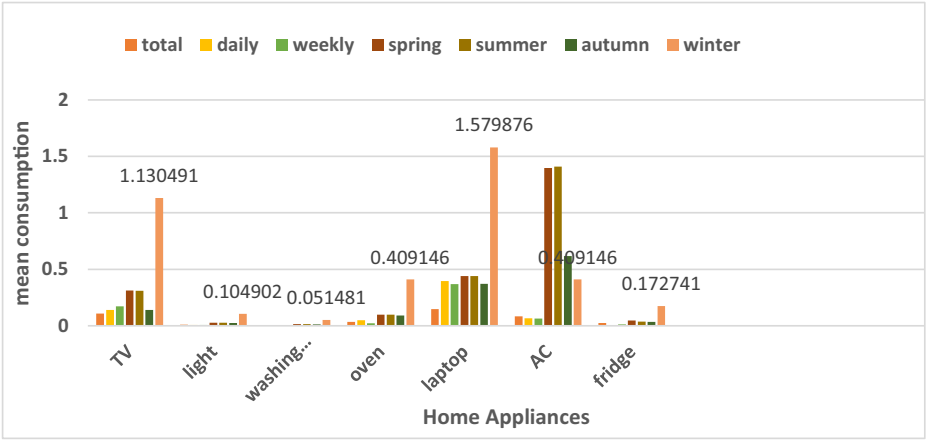

Fig. 8 Mean consumption

$$1*60\ \text{min} = 60\ \text{unit per hour}$$

Then its daily consumption 60 * 24 h = 1440 unit.

Week contain 7 days then weekly consumption 1440 * 7 = 10,080 unit.

Month contain 4 weeks then Monthly consumption is 10,080 * 4 = 40,320 is shown in Fig. 9.

The segment of the total number of predictions that is correct otherwise overall how often the model forecast correctly payer and non-payer.

Accuracy calculation Accuracy = (True positive+ True Negatives) / (True Positives + True Negatives + False Positives + False Negatives).

False Positives (FP):It is the act of wrongly forecasting the positive classes.

False Negatives (FN): It is the act of wrongly forecasting the negative classes.

True Positives (TP): It is the act of correctly forecasting the positive classes.

True Negatives (TN): It is the act of correctly forecasting the negative classes.

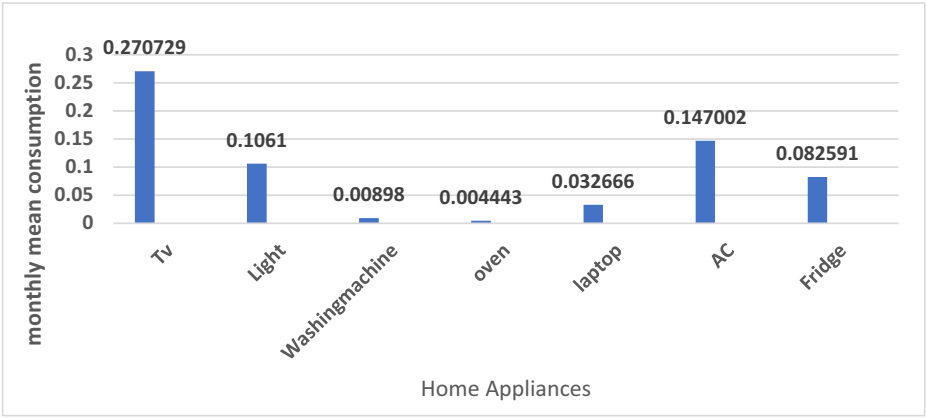

Fig. 9 Monthly consumption

Accuracy is the ratio of rightly forecasted observations to the total observations. It is seem that, if the model has high accuracy then the model is best. It is a significant factor that values of false positive and false negatives are almost equal for symmetric datasets which we have.

Precision It is the ratio of rightly forecasted positive scrutiny to the complete forecasted positive observations.

Precision = True Positives / (True Positives + False Positives).

More precision associate with the less false positive rate. Precision of 0.788 is obtained which is good.

Recall It is the ratio of rightly forecasted positive scrutiny to the all scrutiny in actual class.

Recall = True Positives / (True Positives + False Negatives).

The Weighted average of Recall and Precision is termed as F1 Score. Accordingly, F1 Score consider false positives and false negatives into consideration. Instinctively it is not as easy to understand as accuracy, but F1 is often more convenient than accuracy when an uneven distribution of class is there. If false negatives and false positives have similar cost then accuracy proves to be best. If false negatives and false positives have dissimilar cost then recall and precision have to be considered.

General formula F- Measure = 2True Positives / (2True Positives + False Positives + False Negatives).

F1 Score Formula:

$$\text{F1 Score} = 2*(\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall}).$$

4.2 Comparative analysis

Random Forest uses Ensemble Learning technique and it works based on the bagging algorithm. It combines the output of all the trees which is created on the subset of the data and reduces overfitting issue in decision trees. This in turn decreases the discrepancy and increases the exactness.

Support vector machine resolve only classification issues, usually considers only 2 classes. But Random Forest resolves both classification as well as regression issues, which intrinsically suited for mutlticlass (~10) problem. Support vector machine memory usage will be higher and requires high cost of computation. Random forest balances the error, for unbalanced datasets unlike SVM. It uses a rule based instead of distance calculation so feature scaling is not required. Artificial Neural Network requires more computational cost and works better for huge volume of data

Random Forest performance is not affected by nonlinear parameters disparate curve based algorithms. Random Forest may carry out better than other curve based algorithms as it does not require feature scaling and not affected by non-linearity.

Random forest is the technique of machine learning while neural networks are exclusive to deep learning. Easy to make parallel method, training speed is faster in random forest. On the other hand, to become more precise, the recurrent neural network demands much more data than an individual person's data. If neural networks are employed it becomes more tedious

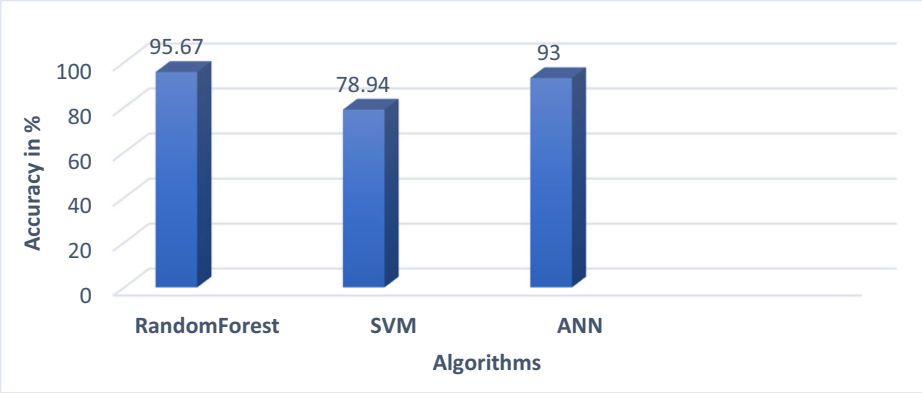

Fig. 10 Accuracy of random forest against SVM and ANN

since we need to know its layers and neuron count in the layer and what activation and initialization should be performed but the Random forest requires less pre-processing and the training process is easier. Random Forest is robust to oddity and can grasp automatically and can judge the importance of the feature and the interaction between different features. Random Forest algorithm is very stable. The overall algorithm is not affected much even if a new data point in dataset is introduced since the new data can influence one tree, but it is very arduous for it to influence all the trees.

The study [18] is anxious with the collation of neural networks and random forest on prophesizing building energy consumption, which is an arithmetical forecasting and not a classification case. According to the study, Random forest performed little efficient than the neural networks as it productively handles any missing values and can exactly forecast even some of the input values were mislaid. It is less affected by noise and is clearly the best classifier as it achieves the best categorization results. The results of the neural network in average worse case is shown in Fig. 10 [8].

SVM algorithm goal is to draw the decision boundary that can separate n-dimensional space into classes such that new data points can be correctly categorized and the created

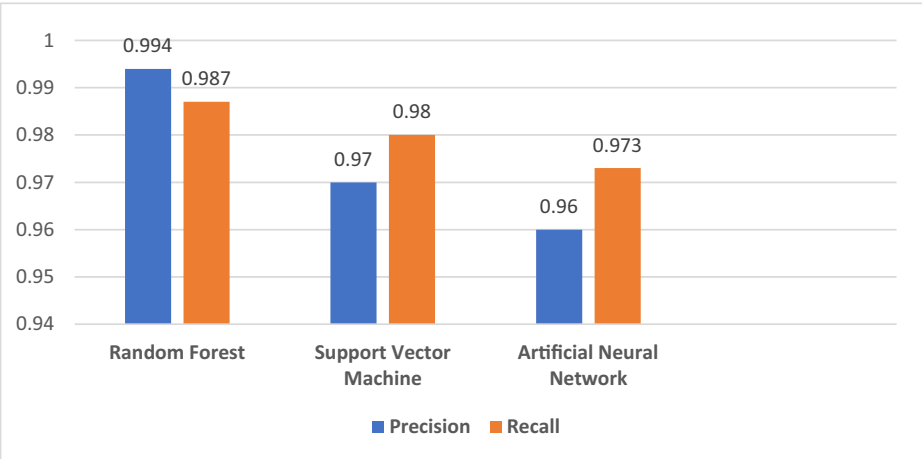

Fig. 11 Precision and recall of random forest against SVM and ANN

decision boundary is also known as hyperplane. So the accuracy of the prediction of SVM is shown as 78.54%. Since Random forest has so many decision trees to accurately predict the accuracy of this algorithm is comparatively high.

The comparative results of the three algorithms with respect to precision and recall has been illustrated in Fig. 11. Since precision represents the ratio of the correctly forecasted positive analysis to the complete forecasted positive observations, it is observed that random forest depicts 0.2% to 0.3% superior precision than SVM and ANN. Similarly it shows 0.1% improved recall than the other two since recall is the ratio of rightly forecasted positive analysis to the all analysis in actual class.

5 Conclusion

The flourish of smart meter paved a way for availability of information about how consumers uses the electrical energy across the country during varied seasons. This work performs the forecasting the consumption of electricity units and analyses the peak demand using efficient machine learning algorithm with the smart meter dataset. The proposed method uses Random forest classification technique to forecast the units and price for the different intervals of time for the various home appliances. Among the various classification models our approach outperforms the other algorithms for a large smart meter dataset with a performance accuracy of 95.67% and improved accuracy of precision and recall based on the obtained results. Since in near future electric vehicles are going to increase in count,the electric consumption of those vehicles can even be included as an additional parameter in the forecasting of electric load and furthermore evolutionary preprocessing tools and algorithms can be employed to improve the prediction accuracy.

References

1. Ahmad MW, Mourshed M, Rezgui Y (2017) Trees vs Neurons: comparison between Random Forest and ANN for high-resolution prediction of building energy consumption. Energy Buildings 147:77–89. https://doi.org/10.1016/j.enbuild.2017.04.038
2. Ali A, Zhu Y,Chen Q, Yu J, Cai H (2019) Leveraging Spatio-Temporal Patterns for Predicting Citywide Traffic Crowd Flows Using Deep Hybrid Neural Networks. 2019 in IEEE Access International Conference on parallel and Distributed Systems. https://doi.org/10.1109/ICPADS47876.2019.00025
3. Canwan, Niu M, Song Y, Xu Z (2017) Pareto Optimal Prediction Intervals of Electricity Price. IEEE 32(1). https://doi.org/10.1109/TPWRS.2016.2550867
4. Chen J, Deng S-J, Huo X (2008) Electricity price curve modelling and forecasting by Manifold learning. Volume: 23. https://doi.org/10.1109/TPWRS.2008.926091
5. Chiaraviglio L, D'andreagiovanni F, Lancellotti R (2018) An approach to balance maintenance costs and electricity consumption in cloud data centres in IEEE Explore. 3. https://doi.org/10.1109/TSUSC.2018.2838338
6. Hamed C, Payam ZD, Palak PP (2018) Electricity price forecasting for operation scheduling of behind the meter storage system. Volume :9. https://doi.org/10.1109/TSG.2017.2717282
7. Chou J-S , Hsu S-C, Ngo N-T, Lin C-W, Tsui C-C (2018) Hybrid Machine Learning System to Forecast Electricity Consumption of Smart Grid-Based Air Conditioners. https://doi.org/10.1109/JSYST.2018.2890524
8. Edomah N (2017) Modelling Future Electricity: Rethinking the organisational model of Nigeria 's Electricity sector in IEEE Explore. 5. https://doi.org/10.1109/ACESS.2017.2769338

9. Fan G-F, Peng L-L, Hong W-C, Sun F (2016) Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.08.051
10. Fernandez-Delgo M, Cernadas E, Barro S (2014) Do we need hundreds of classifiers to solve real world classification problems? J Machine Learn Res 15(15):3133–3131
11. Geetha R, Thilagam T, (2020) A review on the effectiveness of machine learning and deep learning algorithms for cyber security, archives of computational methods in engineering. https://doi.org/10.1007/s11831-020-09478-2
12. Gonzalez AM, Roque AMS, Garcia-gonzalez J. (2005) Modelling and forecasting electricity price with input and output hidden Markov in IEEE Explore 20(1). https://doi.org/10.1109/TPWRS.2004.840412
13. Guo S, Lin Y, Feng N,Song C, Wan H (2019) Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting,2019 in Association for the Advancement of Artificial Intelligence. 33(1). https://doi.org/10.1609/aaai.v33i01.3301922
14. Hobby JD, Shosthishvili A, Tucci GH (2012) Analysis and Methodology to segregate residential elctricity consumption in different taxonomies in IEEE Transactions. 3. https://doi.org/10.1109/TSG.2011.2167353
15. Hong W-C (2011) Electric load forecasting by seasonal recurrent SVR with chaotic artificial bee colony algorithm. Energy 36:5568–5578. https://doi.org/10.1016/j.energy.2011.07.015
16. Hong W-C, Dong Y, Zhang WY, Chen L-Y, Panigrahi BK (2013) Cyclic electric load forecasting by seasonal SVR with chaotic genetic algorithm. Int J Electric Power Energy Syst 44:604–614. https://doi.org/10.1016/j.ijepes.2012.08.010
17. Imtiaz A K, Norman B Mariun, Amran MMR, M Saleem, N I A Wahab and Mohibullah S (2006) Evaluation and Forecasting of Long Term Electricity Consumption Demand for Malaysia by Statistical Analysis" https://doi.org/10.1109/PECON.2006.346658
18. Karimtabar N, Alipour SPS (2015) Analysis and predicting electricity energy consumption using data mining techniques- Acase study I.R. Iran -Mazandaran province in IEEE Explore. https://doi.org/10.1109/PRIA.2015.7161634.
19. Khan I, Huang JZ, Masud Md Ab (2016) Segmentation of factories on electricity consumption behaviour using load profile data. volume 4. https://doi.org/10.1109/ACCESS.2016.2619898
20. Kwac J ,Flora J, Rajagopal R (2018) Lifestyle segmentation based on Electricity consumption days in IEEE Transactions on Smart Grid. 9. https://doi.org/10.1109/TSG.2016.2611600
21. Li M-W, Geng J, Hong W-C, Zhang L-D (2019) Periodogram estimation based on LSSVR-CCPSO compensation for forecasting ship motion. Nonlinear Dynamics 97(4):2579–2594. https://doi.org/10.1007/s11071-019-05149-5
22. Li S, Yang J, Song W, Chen A (2019) A real time electricity scheduling for residential home energy management in IEEE Explore. 6 https://doi.org/10.1109/JIOT.2018.2872463
23. Lin J ,Yu W, Yang X (2016) Towards Multistep Electricity price in sMart grid electricity markets. 27. https://doi.org/10.1109/TPDS.2015,2388479
24. Min TT, Poor HV (2011) Scheduling power consumption with price uncertainty in IEEE Explore. 2. https://doi.org/10.1109/TSG.2011.2159279
25. Molla T, Khan B, Moges B, Alhelou HH, Zamani R, Siano P (2019) Integrated Optimization of Smart Home Appliances with Cost-effective Energy Management System in IEEE Explore. https://doi.org/10.17775/CSEEJPES.2019.00340
26. Ranjbar M, N. Sadati, Soleymani (2006) Electricity Price Forecasting Using Artificial Neural Network in IEEE Explore. https://doi.org/10.1109/PEDES.2006.344294
27. Valentin Robu, Vinyals M, Rogers A, Jennings NR (2018) Efficient buyers groups with predictions of use electricity tariffs in IEEE Explore. 9. https://doi.org/10.1109/TSG.2017.2660580
28. Sharma S, Xu Y, Verma A, Panigrahi BK (2018) Time-Coordinated Multi-Energy Management of Smart Buildings under Uncertainties in IEEE Explore https://doi.org/10.1109/TII.2019.2901120
29. Hideitsu H, Haoyang S, Noboru M, Shinji W, Yasuhiro H (2013) A Versatille clustering method for electricity consumption pattern analysis in household. Volume 4. https://doi.org/10.1016/j.apenergy.2014.08.111
30. Xiang M, Rao H, Tan T, Wang Z, Ma Y (2019) Abnormal behaviour analysis algorithm for electricity consumption based on density clustering in IEEE Explore. 2019. https://doi.org/10.1049/joe.2018.5123
31. Zedan FM, Mohammad AS, Zakhary SZ (2010) A Non zero sum approach to interactive electricity consumption in IEEE. 25(1). https://doi.org/10.1109/TPWRD.2009.2031647
32. Zichen Zhang · Wei-Chiang Hong,"Electric load forecasting by complete ensemble empirical model decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm. Nonlinear Dynamics", 2019, 98, 1107–1136. https://doi.org/10.1007/s11071-019-05252-7

33. Zhang Z, Ding S, Jia W (2019) A hybrid optimization algorithm based on cuckoo search and differential evolution for solving constrained engineering problems. Eng Appl Artificial Intell 85:254–268. https://doi.org/10.1016/j.engappai.2019.06.017
34. Zhang W,Li XDH, Xu J (2020) Unsupervised detection of abnormal electricity consumption behaviour based on future Engineering in IEEE Explore. 8. https://doi.org/10.1109/acess.2020.2980079
35. Zhang Z, Ding S, Sun Y (2020) A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task. Neurocomputing 410:185–201. https://doi.org/10.1016/j.neucom.2020.05.075
36. Zhao JH, Dong ZY, Li X, Wong KP (2007) framework for electricity price spoke analysis with advanced datamining methods in IEEE Transactions on Power systems. 22. https://doi.org/10.1109/TPWRS.2006.889139
37. Zhou S, Zhou L, Mao M, Tai H-M, Wan Y (2019) An optimised heterogeneous structure LSTM network1 for electricity price forecasting in IEEE Explore. 7. https://doi.org/10.1109/ACESS.2019.2932999