Electrical Load Prediction in Energy Internet via Linear Correlation Coefficient Approach

Yangyang Ming, Junwei Cao*
Research Institute of Information Technology,
Tsinghua University
Beijing, China
e-mail: jcao@tsinghua.edu.cn

***Abstract*—Focusing on the prediction of electrical load in different time scales in energy Internet (EI), we propose a linear correlationbased load predicting algorithm which predicts the future load using the linear statistical correlation in load data sequence. The correlation factor is calculated, and its absolute value is used to filter out the relevant sample sequences.Then the candidate values are estimated using a linear fit algorithm.These candidate values are further weighted averaged such that the anticipated load values are obtained. Through preliminary simulations, we can realize an effective forecast of load data.**

Keywords—Energy Internet; linear correlation; load predicting; statistical correlation;linear fit

I. INTRODUCTION

With the development of global economy, the global electric energy consumption is rapidly increasing [1]. From the worldwide point of view, Energy Internet(EI) has been considered as one of the presentative modern energy systems [2]. The concept of EI is based on the high efficiency of information and communication infrastructure [3]. By combining the advantages of Internet of Things (IoT) and Cyber-Physical integration system, and by applying distributed energy utilization and multi-energy complementation related technologies, EI can realize an overall coordination of power resource, electric network, load and energy storage [4].

A rapid and precise forecasting of consumers' electric load has become the pre-condition forrealizing a robust and efficient energy management in modern power systems. Besides, it can improve the energy utilization efficiency, and reduce the energy utilization cost through demand side management [5] and demand response. It is notable that load forecast should be performed in different time and space dimensions. In the field of EI, intelligent control which requires the full information and characteristics of electrical loads has become popular. Relying on wide utilization of intelligent sensor devices and strong support of advanced information and communication infrastructures, precise load prediction isfeasible.

The electric energy forecast has attracted much attention, and advances onvarious types of algorithms are made. In [6] and [7], the forecast types are classified rationally. We classify the related techniques as follows:

1. In [8]-[13], forecasts based on pattern segment and classifications are presented. In [8], the historical data is divided according to the results of the k-means clustering algorithm. In [9] and [10], the forecasting methodsusing support vector machine (SVM)for classification are proposed. In [11], artificial neural networks (ANNs) is applied in the algorithm as the classification tool for the load forecast. Multiple classification algorithms are proposed and tested in [12]. In [13], a structured framework and a discriminative index that can be used to segment the load data along multiple contextual dimensions areintroduced. The classification number becomes the uncertain factor for this type of forecast, which is important and apparently influences the final performance.
2. In [14]-[24], forecasts based on parameter fitting are investigated. The typical technologies include different forms ofANNs[14]-[18], and auto-regression based forecasts considering different types of internal and external variables [19]-[22], and SVM based forecasts which use the SVM algorithm to train related coefficients [23], [24]. Although this type's algorithm is relative simple, more external conditions are required to be considered, such as meteorology and financial related variables which sometimes cannot be obtained.
3. In [25] and [26], forecasts based on probability estimation are proposed. In [25], Bayesian network is used in the predict algorithm. The algorithm proposed in [26] simultaneously uses neural networks and adaptive Bayesian learning. In this case, defining the probability distribution isrelatively difficult.
4. In [27]-[30], forecasts based on intelligence technology are presented. In [27], expert knowledge is used. In [28], the concept of big data is used. In [29],machine learning is used in the forecast. In [30], thefactored conditional restricted Boltzmann machine is used in the forecasting algorithm. The prediction performance of such forecasting type is based on the total quality of the data

5. In [31], forecast based on ensemble technologies is proposed. By using different prediction algorithms at the same time through weighted averaging, there is a high probability of performance improvement.

6. In [32], forecast based on compressive sensing is proposed, which can improve the performance from another perspective.

7. In [33] and [34], forecast based on sequence and block is investigated.Although such forecasting methodwhich can be used for long sequence, the error may be accumulated in the iterations.

In addition, the electrical load forecast can be classified based on the time dimension, such as short-term, medium-term, and long-term [35].

In this paper, our proposed algorithm is partly based on sequence and block related algorithms and the parameter fitting theory.The statistical correlation in the sequence through big data analysis is implemented.We further explore the linear fitting to estimate the load values, which has firm mathematical foundation and embodies the advantages of big data.Our approach can be used in many time dimensions and tested in short time load forecast, which is novel in the field of electric load forecast.

The subsequent framework of this paper is as follows. In Section II, an algorithm for load forecasting is proposed. In Section III, simulation results are illustrated. Finally, Section IV gives a conclusion.

II. Algorithm Illustration

The proposed algorithm is based on the character of linear correlation coefficient which states that if the coefficient value is larger than 0.95, the two sequences have a property of highly linear correlation, and if two sequences'correlation coefficient is equal to 1, then there exists a linear expression between the two sequences with probability one($y = ax + b$ for sequence x and y).

1. We want to predict the load value at time *t*, day *d*(called expected value). We first fetch the load sequence just before time *t* in day *d* with length n(called original sequence, denoted as $\mathbf{x_n} = x_1, x_2, ...x_n$). We then select the reference sequences with the same length from the interval window $[t-4, t+1]$(elements in this window denoted as the end point of the sequence with length *n*), which arerecorded before day *d*(called reference sequence,denoted as $\mathbf{y_n} = y_1, y_2, ...y_n$).

2. For each reference sequence, we calculate the correlation coefficient between itself and itscorresponding original sequence. As most data has some statisticalsimilarity for each day's value especially near the same time point, there will comprise a large part of sequences'correlation coefficients whose value are near or equal to 1.

$$Cn = corr(\mathbf{x_n}, \mathbf{y_n}) =$$

$$\sum((x_n - mean(\mathbf{x_n})) * (y_n - mean(\mathbf{y_n}))) / ...$$

$$\sqrt{\sum(x_n - mean(\mathbf{x_n}))^2 * \sum(y_n - mean(\mathbf{y_n}))^2}$$

3. From the correlation theory,when the correlation coefficient of two data sequences is near to 1, there will be a large linear relation, such as

$$y_1 = ax_1 + b$$

$$y_2 = ax_2 + b$$

...

$$y_n = ax_n + b$$

Here, we calculate the linear coefficient between each reference sequence and its corresponding original sequenceusing linear fit methodwhen correlation coefficient nears to 1, and get the corresponding parameter value as *a* and *b*.

4. As the results shown in the simulation part, we can assume that the linear correlation value will not change too much, if we extend the sequence length with one more data. We use the load value immediately after the reference sequence, combined withcalculated linear coefficients; we estimate the expected value corresponding to each reference sequence as each instance:

$$\bar{y} = ax + b$$

5. We weighted average all of the estimated instance values, and set the weight based on the Euclideandistance between the reference sequence and original sequence.

$$weight = mse(\mathbf{x_n} - \mathbf{y_n}) / \max(mse(\mathbf{x_n} - \mathbf{y_n}))$$

$$weight = \exp(Cn - weight)$$

$$\tilde{y}_{n+1} = \sum weight * \bar{y}_{n+1} / \sum weight$$

$\bar{y}_{n+1}$ is the candidate estimate value of y_{n+1}.

And as a comparison, the mean result is obtained as follows.

$$\bar{\bar{y}}_{n+1} = \sum \bar{y}_{n+1} / \left\| \bar{y}_{n+1} \right\|_0$$

6. We comparethe estimated result of above steps with the expected value and calculate the relative error,

$$MAPE1 = abs(\tilde{y}_{n+1} - y_{n+1}) / abs(y_{n+1})$$

$$MAPE2 = abs(\bar{\bar{y}}_{n+1} - y_{n+1}) / abs(y_{n+1})$$

7. We run this algorithm on 100 days and get 100 *MAPE* values, which is further averaged to get the final estimated error performance.

$$MAPE = sum(MAPEn) / count(MAPEn)$$

Through *MAPE* value, we can decide the optimal parameter value of this algorithm for different scenes.

III. Simulation Result and Analysis

As a preliminary research, we first choose the data of one meter to test this algorithm. We find that the data vary stochastically, and the correlation algorithm cannotlead to good results (as we found, other classical algorithms, such as ANN and linear regression,cannotperformwell in this situation, either). Then, we combine the metering data of 8

meters in the same region, which could get comparable results with ANN.

The data in each meter is sampled in each half hour with 7*48 recordseach week for 537 continuous days; we use the latest 100 days' data for algorithm testing, with all the data before the test days as their history data(for model training).

We first calculate the distribution of correlation coefficients'differences.The detailed algorithm is as follows:

1. We select the data sequences with length 6 takenin step1 of Section II asoriginal sequence and reference sequence and calculate the correlation coefficient betweeneach pair of sequence (denoted as co_1).
2. If the correlation coefficient is larger than 0.95, we extend each candidate sequence by one more element and also calculate the correlation coefficient, (denoted as co_2).
3. We calculate the difference of co_1 and co_2 , which is $delta\_co = co\_1 - co\_2$.
4. We ergodic through all the sequencenear the same time point, forming the histogram for $delta\_co$.

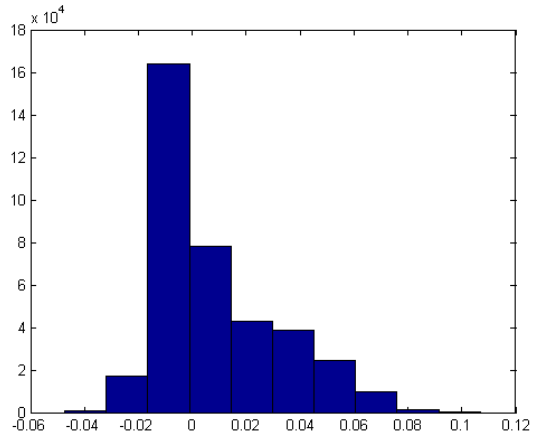

Figure 1.The result of 10-1.

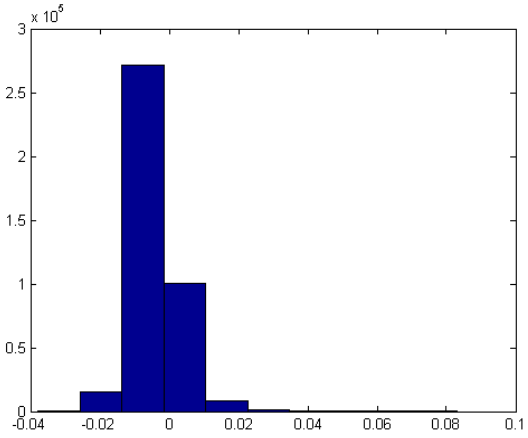

Figure2.The result of 20-1.

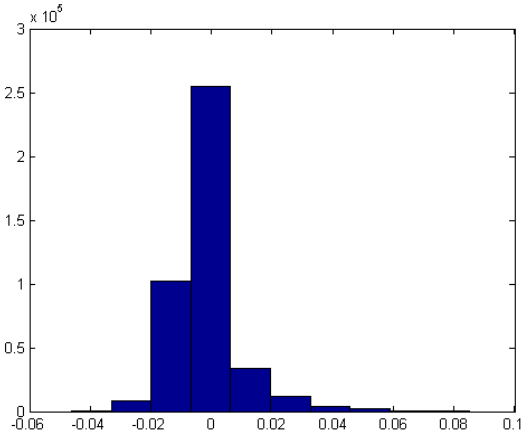

Figure3.The result of 30-1.

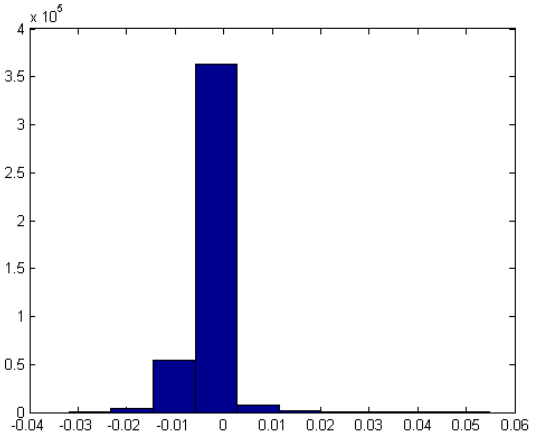

Figure4.The result of 40-1.

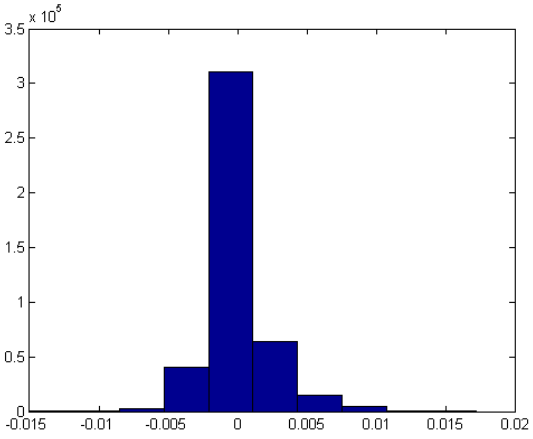

Figure5.The result of 10-8.

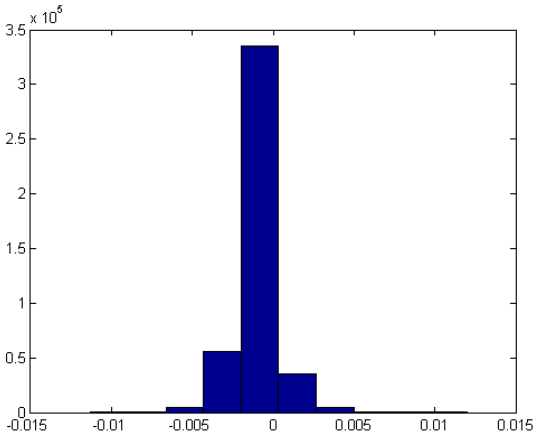

Figure6.The result of 20-8.

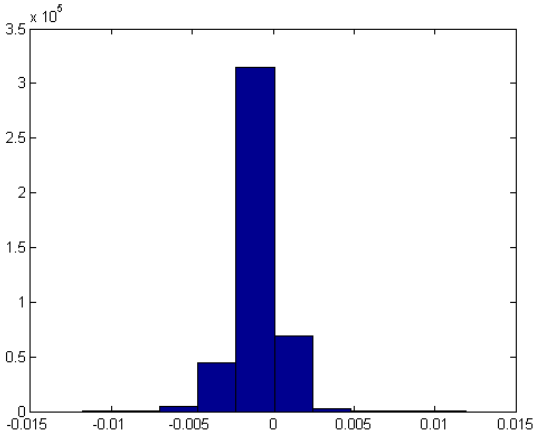

Figure7.The result of 30-8.

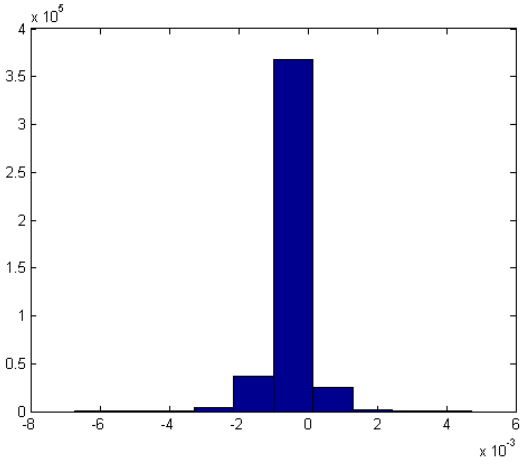

Figure8.The result of 40-8.

5. The algorithm can be easily extended to generalizedscenarios, such as united showing of different correlation coefficients and different lengths.

The related histogram results are shown from Fig.1 to Fig. 8.Characters *m-n* in the caption of those figures determine the result oftotal n meters at the time position m (from 0 to 48 for hour 00:30 to 24:00 in one day).

From the above figures, we can see that when the histogram is skewed to the left part,the relevant linear correlation between two sequences becomes higher, and the performance of the algorithm will be better, which is proved by subsequent experiments.

As the above figures show, the relevant correlationperformance with 8 meters is apparently higher than 1meter, and the sorted algorithm performances in 8 meters from low to highare: position 10, position 30, position 20, and position 40 accordingly. This trend can be easily seen in the above figures and verified by simulations.

The simulation resultsare listed in TABLE I, in comparisonwiththe results obtained using ANN algorithm.In this table, the combined algorithm uses $result = (canidate1 + candidate2)/2$ as the final result.

The reason forthe performancebeing not high is mainly due to the characters of the test data, which is metered very locally andsporadically (in small space scale).Although we further use the sum of 8 meters, the problem is only alleviated in medium extent. And when more macro scale data is obtained, the simulation performance can still be improved, which will promote the further research.

From the result of TABLEI, we can see that the performance of the proposed algorithm is comparable to the ANN algorithm, both of whose complexity are not very high.

When increasing the candidate sample number, the performance is improved from 8 to 12, and then to 16, there is no further improvement from 16 to 18, which means the samples are saturated at near 18 samples.

When we compare the result of meanbased algorithm and the weighted average based algorithm, we can easily see that in most conditions, weighted average method is better than mean method, which is coincide with the intuition of the people. And the median has higher performance in some simulation instances as compared to mean.

When examine the result of the combined algorithms, their performance is not always improved compared to the corresponding single result, but some extra good value can be seen from the table, which embodies the potential value of combining other multiple algorithms with our proposed method.

IV. CONCLUSION

In this paper, we propose a simple electrical load predict algorithm based on correlation coefficient. Due to its low complexity, it can be used in many time scales, especially the short time scale, and can be further jointly used with other algorithms such as ANN and linear regression. And from the results, we can see that its performance is comparable with ANN. If the relevant character of linear correlation is improved, the simulation result will be improved as well.

TABLE I.SIMULATION RESULT：*MAPE*VALUE

Position 10	Position 20	Position 30	Position 40	Detailed explanation
0.3213	0.2166	0.2298	0.1541	ANN length=6
0.3470	0.2199	0.2411	0.1677	ANN length=8
0.3432	0.2436	0.2316	0.1544	ANN length=10
0.3463	0.2162	0.2451	0.1654	Tested algorithm with median result
0.3401	0.2133	0.2427	0.1711	Tested algorithm with mean result, correlation coefficient>0.9999
0.3332	0.2469	0.2478	0.1660	Tested algorithm with mean result, candidate sample=8
0.3342	0.2369	0.2427	0.1640	Mean, candidate sample=12
0.3342	0.2322	0.2378	0.1603	Mean, candidate sample=16
0.3343	0.2304	0.2380	0.1610	Mean, candidate sample=18
0.3321	0.2437	0.2433	0.1670	Weighted average, candidate sample=8
0.3334	0.2355	0.2405	0.1641	Weighted average, candidate sample=12
0.3341	0.2317	0.2350	0.1605	Weighted average, candidate sample=16
0.3339	0.2300	0.2360	0.1613	Weighted average, candidate sample=18
0.3341	**0.2117**	0.2387	0.1583	Combined correlation coefficient>0.9999,mean with ANN(length=8),average with the two results
0.3321	0.2253	0.2367	**0.1540**	Combined correlation sample=16,mean with ANN(length=8),average with the two results
0.3284	0.2280	0.2319	0.1547	Combined correlation sample=12,mean with ANN(length=8),average with the two results
0.3302	0.2295	**0.2287**	0.1544	Combined correlation sample=16,mean with ANN(length=6),average with the two results

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (grant No. 61472200) and Beijing Municipal Science & Technology Commission (grant No. Z161100000416004).

REFERENCES

[1] G.Venkataramanan and C. Marnay, "A larger role for microgrids,"*IEEEPower Energy Mag.*, vol. 6, no. 3, pp. 78-82, May 2008.

[2] J. Cao and M. Yang, "Energy Internet - towards smart grid 2.0,"in *Proc.Fourth Int.Conf. Networking & Distributed Computing*, Los Angeles, USA, Dec. 2013, pp. 105-110.

[3] J. Cao,M. Yang, and D. Zhang, "Energy internet: an infrastructure for cyber-energy integration," *Southern Power System Technology*,vol.8, pp. 1-10, 2014.

[4] H. Hua, J. Cao, G. Yang, and G. Ren, "Voltage control for uncertain stochastic nonlinear system with application to energy Internet: Non-fragile robust H_∞ approach," *J. Math. Anal. Appl.*, vol. 463, no. 1, pp. 93-110, 2018.

[5] X. Chen, Y. Zhou, W. Duan, J. Tang, and Y. Guo, "Design of intelligent demand side management system respond to varieties of factors," in *Proc. China Int.Conf. Electricity Distribution*,Nanjing, China,2010, pp. 1-5.

[6] H.K. Alfares and M. Nazeeruddin, "Electric load forecasting: Literature survey and classification of methods," *Int. J. Syst. Science*, vol.33, pp.23-34, 2002.

[7] P. Mirowski, S. Chen, T.K. Ho,and C.N. Yu, "Demand forecasting in Smart Grids," *Bell Labs Tech. J.*, vol. 18, pp.135-158, 2014.

[8] N. Arghira, L. Hawarah, S. Ploix, and M. Jacomino, "Prediction of appliances energy use in smart homes," *Energy*, vol. 48, pp. 128-134, 2012.

[9] A. Jain and B. Satish, "Clustering based short term load forecasting using support vector machines," in *Proc. IEEE Bucharest PowerTech*, Bucharest, Romania, 2009,pp.1-7.

[10] B.J. Chen, M.W. Chang, and C.J. Lin, "Load forecasting using support vector machines: A study on EUNITE Competition 2001," *IEEE Trans. Power Syst.*, vol. 19, pp. 1821-1830, 2004.

[11] F.D.O.Saraiva, W.M.S. Bernardes, and E.N. Asada, "A framework for classification of non-linear loads in smart grids using artificial neural networks and multi-agent systems," *Neurocomputing*, vol. 170, pp. 328-338, 2015.

[12] C.Y. Kuo, M.F. Lee, C.L. Fu, Y.H. Ho, and L.J. Chen,"An in-depth study of forecasting household electricity demand using realistic datasets," in *Proc. Int. Conf. on Future Energy Syst*, 2014, pp. 145-155.

[13] T.K. Wijaya,T. Ganu,D. Chakraborty,K. Aberer,and D.P. Seetharam, "Consumer segmentation and knowledge extraction from smart meter and survey data," in *Proc. Siam Int. Conf. on Data Mining*, 2014, total page:9.

[14] A.G. Bakirtzis, V. Petridis, S.J. Kiartzis, M.C. Alexiadis, and A.H. Maissis, "A neural network short term load forecasting model for the Greek power system," *IEEE Trans. Power Syst.*, vol. 11, pp. 858-863, 1996.

[15] D.C. Park, M.A. El-Sharkawi, R.J.Marks, L.E. Atlas and M.J. Damborg, "Electric load forecasting using an artificial neural metwork," *IEEE Trans. Power Syst.*,vol. 6, no. 2, pp. 442-449, May 1991.

[16] A. KAotanzad, R. Afkhami-Rohani, and D. Maratukulam, "ANNSTLF - artificial neural network short-term load forecaster -Generation three," *IEEE Trans. Power Syst.*, vol. 13, pp.1413-1422, Nov. 1998,

[17] A. Tascikaraoglu , O. Erdinc, M. Uzunoglu, and A. Karakas, "An adaptive load dispatching and forecasting strategy for a virtual power plant including renewable energy conversion units,"*Appl. Energy*, vol. 119, pp.445-453, 2014.

[18] K. Bhaskar and S.N. Singh, "AWNN-assisted wind power forecasting using feed-forward neural network," IEEE Trans. Sustain. Energy," vol.3, no. 2, pp.306-315, Apr. 2012.

[19] S. Vemuri, W.L. Huang and D.J. Nelson, "On-Line algorithms for forecasting hourly loads of an electric utility," *IEEE Trans. Power Apparatus and Syst.*, vol. PAS-100, pp. 3775-3784, Aug. 1981.

[20] Y. Dong, J. Wang, H. Jiang, and J. Wu, "Short-term electricity price forecast based on the improved hybrid model," *Energy Conversion & Management*,vol. 52, pp. 2987–2995,2011.

[21] M.R. Braun, H. Altan, and S.B.M. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," *Appl. Energy*, vol. 130, pp. 305–313, 2014.

[22] J.C. Lam, K. K.W. Wan, D. Liu, and C.L. Tsang, "Multiple regression models for energy use in air-conditioned office buildings in different climates," *Energy Conversion & Management* , vol. 51, pp. 2692-2697, 2010 .

[23] C. Chen and D.J. Cook, "Behavior-based home energy prediction, " in *Proc. 8thInt. Conf. on Intelligent Environments*, Guanajuato, Mexico, Jun. 2012, pp. 57-63.

[24] M. Saadat and S. Ghili, "Electrical energy modeling in Y2E2 building based on distributed sensors information," pdfs.semanticscholar.org.

[25] L. Hawarah and M. Jacomino, "User behavior prediction in energy consumption in housing using Bayesian networks," in *Proc. Int. Conf. Artificial Intelligence & Soft Computing*, 2010 , pp.372-379.

[26] R. Blonbou, "Very short-term wind power forecasting with neural networks and adaptive Bayesian learning," *Renew. Energy*, vol. 36, pp. 1118-1124, 2011.

[27] K. Basu, L. Hawarah, N. Arghira, H. Joumaa, and S. Ploix, "A prediction system for home appliance usage," *Energy and Buildings*, vol. 67, pp. 668–679, 2013.

[28] 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, 2016.

[29] R.E. Edwards, J.New, and L.E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.

[30] E. Mocanu, P.H. Nguyen,M. Gibescu,E.M. Larsen,and P. Pinson, " Demand forecasting at Low aggregation levels using factored conditional restricted Boltzmann machine," in *Proc. Power Systems Computation Conf.*, Genoa, Italy, Jun. 2016, pp.1-7.

[31] S. Wen, V. Babushkin, Z. Aung, and L.W. Wei, "An ensemble model for day-ahead electricity demand time series forecasting," in *Proc. Int. Conf. on Future Energy Systems*, 2013, pp.51-62.

[32] A. Tascikaraoglu and B.M. Sanandaji, "Short-term residential electric load forecasting: A compressive spatio-temporal approach," *Energy and Buildings*, vol. 111, pp. 380-392, 2016.

[33] F.M. Alvarez, A. Troncoso, J.C. Riquelme, and J.S.A. Ruiz, "Energy time series forecasting based on pattern sequence similarity," *IEEE Trans. on Knowledge and Data Engineering*, vol. 23, pp. 1230–1243, 2011.

[34] F. Yu and Y. Hayashi, "Pattern sequence-based energy demand forecast using photovoltaic energy records," in *Proc. Int. Conf. on Renewable Energy Research & Applications*, 2012, pp. 1-6.

[35] E.A. Feinberg,D. Genethliou, "Load forecasting," *Appl. Mathematics for Restructured Electric Power Syst.*," Springer, US, 2005, pp. 269–285.