Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6610273, 11 pages
https://doi.org/10.1155/2021/6610273

Research Article

Appling an Improved Method Based on ARIMA Model to Predict the Short-Term Electricity Consumption Transmitted by the Internet of Things (IoT)

Ni Guo[1], **Wei Chen**[1,2], **Manli Wang**[3], **Zijian Tian**[1] and **Haoyue Jin**[1]

[1]*School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China*
[2]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China*
[3]*School of Physics &Electronic Information Engineering, HeNan Polytechnic University, Jiaozuo 454000, China*

Correspondence should be addressed to Wei Chen; chenwdavior@163.com

Received 24 December 2020; Revised 1 March 2021; Accepted 29 March 2021; Published 12 April 2021

Academic Editor: Chi-Hua Chen

The rapid development of the Internet of Things (IoT) has brought a data explosion and a new set of challenges. It has been an emergency to construct a more robust and precise model to predict the electricity consumption data collected from the Internet of Things (IoT). Accurately forecasting the electricity consumption is a crucial technology for the planning of the energy resource which could lead to remarkable conservation of the building electricity consumption. This paper is focused on the electricity consumption forecasting of an office building with a small-scale dataset, and 117 daily electricity consumption of the building are involved in the dataset, among which 89 values are selected as the training dataset and the remaining 28 values as the testing dataset. The hybrid model ARIMA (autoregression integrated moving average)-SVR (support vector regression) is proposed to predict the electricity consumption with different prediction horizons ranging from 1 day to 28 days. The model performances are assessed by three evaluation indicators, respectively, are the mean squared error (MSE), the root mean square error (RMSE), and the mean absolute percentage error (MAPE). The proposed model ARIMA-SVR is compared with the other four models, respectively, are the ARIMA, ARIMA-GBR (gradient boosting regression), LSTM (long short-term memory), and GRU (gated recurrent unit) models. The experiment result shows that the ARIMA-SVR model has lower prediction errors when the prediction horizon is within 20 days, and the ARIMA model is better when the prediction horizon is in the interval of 20 to 28 days. The provided method ARIMA-SVR has higher flexibility, and it is a great choice for electricity consumption prediction with more accurate results.

1. Introduction

Nowadays, with the continuous increase of the electricity consumption in buildings, the problem of excessive waste of resources has occurred. Internet of things (IoT) has been popularly applied to smart city controls for collecting electricity consumption; specifically, the real-time electricity consumption data is transmitted to the electricity consumption monitoring system by the distributed wireless sensor network (WSN). The wireless sensor is mainly composed of many intelligent distributed wireless sensor nodes, and each of which has the function of sending a message. Forecasting electricity demand in advance and scheduling accordingly is an essential measure to achieve energy conservation, and it could also assist policymakers and energy managers to make reasonable strategies to promote environmental protection and reduce carbon emission.

Electricity consumption is influenced by many factors such as weather conditions, occupant behavior, and the physical parameters of buildings; this could be verified by many relevant publications. Besides, these impact factors are also considered in the experimental analysis as input vectors by researchers. Yannan et al. proposed a framework for data-driven occupant-behavior analytics in Ref. [1] which will

help build an analytics feedback loop from behavior impact to incentive design for energy saving: the approach is designed based on machine learning techniques such as k-means and kernel ridge regression. Climate factors as a type of important factor are analyzed frequently by many researchers in the forecasting field. Caro et al. have studied the temperature's influence and the effect of public holidays on short-term electric load forecasting for Spanish insular electric systems in Ref. [2]; a mathematical 24 h Reg-ARIMA model as a proposed algorithm is utilized to predict the electric load demand of the next day for ten insular systems located in the two Spanish archipelagos. Although it could improve the precision of the electricity forecast when the impact factors are involved into the model, there are still many limitations when collecting data on influencing factors. Therefore, it is an efficient measure to address the challenge of electricity consumption forecasting by extracting data features from the historical time series.

For univariate time series forecasting, the typical ARIMA algorithm is commonly utilized in many works as a result of requiring very few assumptions in model training and the flexibility of application [3–5]. As mentioned by Jamil [6], the hydroelectricity consumption of Pakistan is predicted up to the year 2030 by using the ARIMA model, and exhaustive statistical analysis and validations have been performed in this research; moreover, a sensitivity analysis is also conducted to study the relation of hydroelectricity consumption to the annual population and GDP growth rate of the country. Li et al. have applied four models to forecast the carbon emission intensity in 2030; the experiment result shows that the ARIMA model is the best-fitted model compared with the other three models [7].

Following many studies in the forecasting field, it can be indicated that satisfactory results could not be attained just with the individual models for all situations. For instance, the characteristics of the time series data may not be captured adequately by individual ARIMA model due to the fact that both the linear and nonlinear features existed in historical data, while the ARIMA model just experts in learning the linear trend of the sample data. Artificial neural networks (ANN) perform well only with sufficient information relying on a large number of historical data. Abundant studies have indicated that the combined methodologies are an advantage of solving the complicated problems concerning time series forecasting owing to the method could benefit from each composition algorithm. In general, the hybrid model has always contained a linear model and a nonlinear model component, and the typical ARIMA is frequently used as a linear model component owing to its advantage of capturing the line characters existing in the data. With regard to the nonlinear model component in the hybrid model, there are many choices available such as machine learning algorithms and statistical methodologies. Machine learning methods that support vector regression (SVR) and gradient boosting regression (GBR) are selected to combine the ARIMA model due to these models having shown great potential in dealing with nonlinear patterns using a small dataset.

The main contributions of this study are demonstrated as follows:

(1) The proposed model ARIMA-SVR can be developed using a small training set while maintaining high accuracy, and few studies are found in electricity consumption prediction using the ARIMA-SVR hybrid model

(2) The proposed model does not need any additional variables just based on a value of its historic observation, and the model is very flexible and explanatory; very few parameters need to be tuned and the model is easy to be implemented

(3) The proposed model combines the advantages of the ARIMA and SVR models, and the nonlinear and linear characters could be well extracted by this hybrid model

This paper is organized as follows: Section 2 has exhibited a comprehensive overview of some notable findings related to this work recently, and the literature reviews are illustrated in different aspects. Section 3 is devoted to introducing the related methodologies which appeared in this paper. Section 4 has presented the procedure of data preprocessing and the construction procedure of prediction models. Section 5 has assessed the simulation results utilizing three evaluation indicators and discussed the performances of prediction models. Finally, the conclusion of the paper is provided in the last section.

2. Literature Review

The improvement of forecasting technology has become a focus in researches towards obtaining a higher accuracy of electricity consumption prediction. This section has presented a comprehensive overview of some superior combined methodologies, and the summary is demonstrated from different aspects, such as the perspective of long term and short term. For instance, many authors have demonstrated the prediction situations using specific cases in different forecast horizons. Kaytez has come up with a hybrid model based on the ARIMA model and LSSVR (least-square support vector machine) and applied it to conduct long-term forecasting of net electricity consumption for Turkey until 2022; the final results demonstrate that the hybrid model ARIMA-LSSVM can generate more realistic and reliable forecasts [8]. With the aim to implement the 24 h-ahead forecasting of the district heat demand of buildings, Eseye and Lehtonen has tested several ML (machine learning) approaches and the excellent result is obtained by the integrated model, namely, EMD-ICA-SVM; it has achieved outperformed forecasting accuracy enhancement compared to the other nine evaluated models [9]. The paper [10] has introduced two novels deep supervised machine learning models including RFEM-GKR (Gaussian Kernel regression model with random feature expansion) model and NPK-NNM (nonparametric based k-NN) model for large-scale utilities and buildings' short- and medium-term load requirement forecasts; the hybrid method RFEM-GKR has remarkable predictor improvements and is proved superior with its high accuracy and stability; the proposed model can be taken as a successful tool to predict energy consumption. The study [11] is focused on the

ultrashort-term (15-minute) predictions of residential electricity of consumption by developing a hybrid model which is based on the Holt-Winters (HW) method and extreme learning machine (ELM) network; the single-model Holt-Winters (HW), extreme learning machine (ELM) network, and long short-term memory network were also established in this research; the experiment result has shown that the proposed HW-ELM model offers more outstanding performance compared with other relevant models.

In addition, some other up-to-date publications about the combined models are also exhibited here. Tascikaraoglu and Uzunoglu have contributed a comprehensive review about wind power forecasting which has outlined various combined forecasting approaches and an up-to-date annotated bibliography of the wind forecasting literature [12]. In Ref. [13], the authors propose a hybrid method based on the combination of autoregressive integrated moving average (ARIMA), artificial neural network (ANN), and the proposed support vector regression (SVR) technique to forecast the yearly peak load and total energy demand of Iran National Electric Energy System; the parameters of the SVR technique are optimized using a particle swarm optimization (PSO) method. Nepal et al. have combined the clustering and the ARIMA model toward electricity load forecasting,; the result has proved that the proposed approach has provided improved accuracy as well as superior performances than that using the ARIMA model alone [14]. The combined method which consists of the ARIMA and NGM methods, namely, the NGM-ARIMA model has been put forward by Ma et al. aimed at accurately predicting South Africa's energy consumption in 2017-2030 [15]; the highest prediction accuracy was achieved by the NGM-ARIMA model, and the prediction result is more close to the actual energy consumption compared to the single ARIMA and NGM model. Gulay and Duru have combined three different models: ARDL (autoregressive distributed lag model), EMD (empirical mode decomposition), and ANN (artificial neural network) for the predictive analytics of energy systems and prices; the proposed hybrid forecasting algorithms provided better results by improving the forecasting accuracy [16]. The publication [17] has displayed a novel hybrid model ANFIS which consolidates both ANN and fuzzy frameworks for prediction future power utilization; the result has proved that this hybridizing approach has the potential of improving prediction performance since it has more significant accuracy and leads to smaller errors contrasted with other models. Likewise, the advantages of the hybrid approach were also verified by many studies [18–20]。.

Different hybrid methodologies are used in different kinds of literature, and the desired results are obtained in various forecasting fields. Generally speaking, it can be concluded that regarding methodologies, the hybrid models or the combined models are composed of the linear and nonlinear models, and each of which carries out a part in the process of the prediction. In addition, after reviewing many related publications in this field, it can be indicated that the hybrid model has a significant advantage of capturing the hidden linear and nonlinear components which are embedded in the original dataset.

3. Methodology

3.1. Autoregressive Integrated Moving Average (ARIMA) Algorithm. The typical ARIMA (autoregressive integrated moving average) algorithm has been proved to be an efficient and reliable method for dealing with the univariable time series. The emphasized advantage is that the ARIMA algorithm does not need any additional variables just based on the values of its historic observations. And the required conditions previous to conduct the ARIMA model process should be satisfied with two conditions; one is that it should be a stationary time series, and the other is the recommended minimum amount of the sample data is at least 50 [21].

Actually, the ARIMA algorithm was integrated by autoregression (AR) and moving average (MA) method with an addition of integrative module; it is characterized by three terms, respectively, p, d, and q, and the general format of the model is ARIMA(p, d, q). Here, p is the order of the AR term, q is the order of the MA term, and d is the number of differencing required for obtaining a stationary time series.

The forecasting equation of the ARIMA (p, d, q) can be expressed as follows:

$$y_t = c + \sum_{i=1}^{p} \varnothing_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t. \tag{1}$$

In the above equation, c is the constant representing the intercept, $\varnothing_i$ and y_{t-i}, respectively, are the parameters and regressors for the AR part of the model, while θ_j and ε_{t-j}, respectively, represent the parameters and regressors of the MA part of the model, whereas ε_t is the white noise error term of the model.

3.2. Support Vector Regression (SVR) Algorithm. SVR (support vector regression) is a good choice to characterize the nonlinear statistical features which existed in the small-scale dataset. This algorithm was firstly proposed by Vapnik et al. in literature [22] and was frequently applied by many researchers in recent years [23, 24]. The fundamental principle of the model is mapping the input data into a high-dimensional space to explore the nonlinear relationship between the input data and output variables; the input dataset is assumed as $\{((x_1, y_1) \cdots (x_n, y_n))\}$, and the optimization is described by the following formula:

$$\min \frac{1}{2} w^T w + C \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right), \tag{2}$$

$$w^T \varnothing(x_i) + b - y_i \le \varepsilon + \xi_i, \tag{3}$$

$$y_i - w^T \varnothing(x_i) - b \le \varepsilon + \xi_i^*, \quad \xi_i, \xi_i^* \ge 0, i = 1, \cdots.., n, \tag{4}$$

where w, b, ξ_i, and ξ_i^* are the decision variable parameters of the optimization problem; $w^T w$ is a regularized term, and ξ_i and ξ_i^* are the slack variables; C is the penalty parameter, ε is the insensitive loss coefficient, it represents a ε tube, if the predicted value is within the tube, the loss is zero, while if it

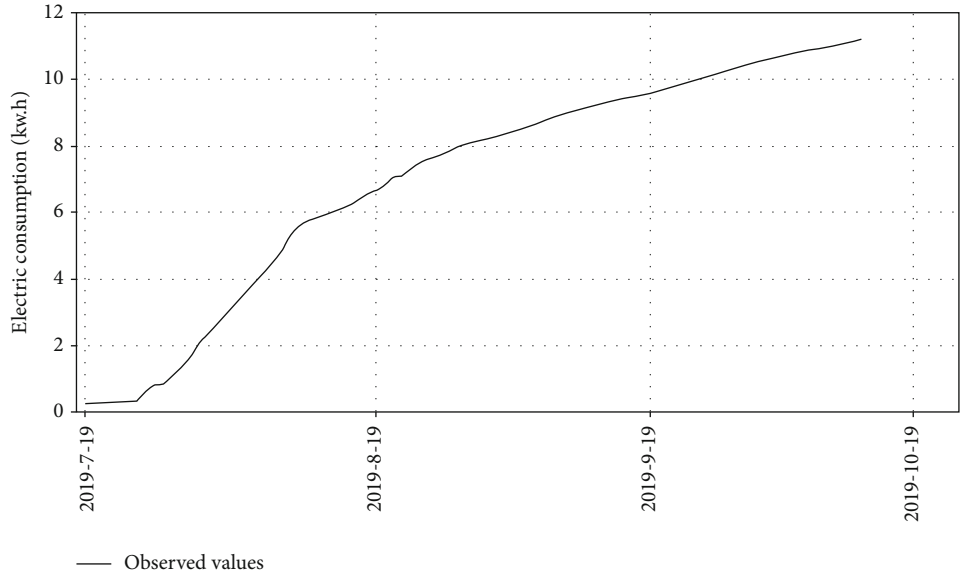


Figure 1: The curve of the original time series.

is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube. What is more, the term $\varnothing(x_i)$ is the feature map and the radial basis function(RBF) is used as the kernel function in this paper; the expression was written as $K(x_i, x_j) = \varnothing(x_i)^T \varnothing(x_i)$.

3.3. Gradient Boosting Regression (GBR) Algorithm. GBR (gradient boosting regression) algorithm is proven as an efficient forecasting technique to capture the nonlinear relationship between the input and output datasets in previous studies. It is one of the boosting regression which has combined a bunch of weak learners [25], and the weak learner would be increasing to minimize the forecasting errors by iteratively training. The main mathematical principles of the gradient boosting regression (GBR) are detailed in the literature [26].

4. Predict Models

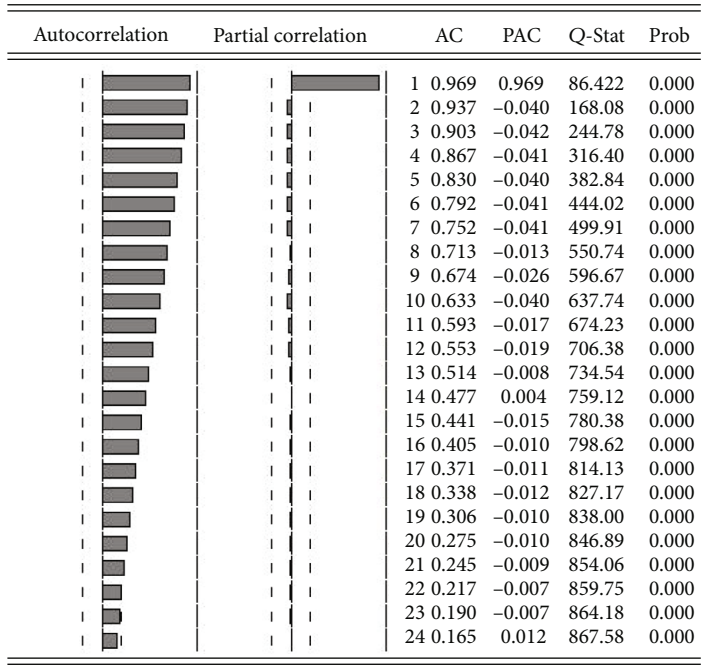
4.1. Data Preprocessing. In many studies, it can be observed that better results are usually obtained when with data preprocessing, and this procedure could make a considerable contribution to the prediction performance of the model. There are two main kinds of processing for the original dataset as the following illustration.

4.1.1. Fill the Missing Data and Normalization. The data is collected from the energy consumption monitoring system by the Internet of Things (IoT), and the monitoring electricity consumption belongs to an office building which is located in Changyang Peninsula in Fangshan District of Beijing, China. The dataset involved 117 daily electricity consumption of the building, and among which, 89 values are selected as the training dataset and the remaining 28 values as the testing dataset. The collected dataset was a series of successive and univariate time series over the period of 19 July 2019 to 12 November 2019. The training dataset was ranged from 19 July 2019 to 15 October 2019, and the testing dataset was the following 28 days. It is an essential step to complete the information of the samples by means of filling the missing data, and the lack information of the data in this paper is filled by calculating the median value of the former and the latter in samples which would be considered to replace the missing data so as to improve the efficiency of the forecasting model. Additionally, data normalization could help reduce the influence of different magnitude levels on the predicted results. The normalization is completed rely on the mapping function which is described as follows:

$$x' = \frac{x - \min}{\max - \min}, \tag{5}$$

where x is the original value of the data and x' is the normalized data through the transform function above, and min and max which appeared in the formula, respectively, represent the maximum and minimum values of the located column in the data. The data could be ultimately mapped to the interval which ranges from 0 to 1.

4.1.2. Differencing. Since the statistical modeling methods assume or require the time series to be stationary, which makes the model easier and more effective, it is necessary to check the stationarity of the time series before it is put into the ARIMA model. The initial time series fluctuation is exhibited in Figure 1; it is clear from the plot that the time series is not stationary which means that it contains the trend and seasonal components. In order to verify the stationarity of the initial data, the detection measures were using the autocorrelation function (ACF) and partial autocorrelation (PACF), and the graph was presented in Figure 2. By observing the values of the autocorrelation in Figure 2, it decays very slowly to zero and it indicates that the initial time series data is not stationary. The procedure



	AC	PAC	Q-Stat	Prob
1	0.969	0.969	86.422	0.000
2	0.937	−0.040	168.08	0.000
3	0.903	−0.042	244.78	0.000
4	0.867	−0.041	316.40	0.000
5	0.830	−0.040	382.84	0.000
6	0.792	−0.041	444.02	0.000
7	0.752	−0.041	499.91	0.000
8	0.713	−0.013	550.74	0.000
9	0.674	−0.026	596.67	0.000
10	0.633	−0.040	637.74	0.000
11	0.593	−0.017	674.23	0.000
12	0.553	−0.019	706.38	0.000
13	0.514	−0.008	734.54	0.000
14	0.477	0.004	759.12	0.000
15	0.441	−0.015	780.38	0.000
16	0.405	−0.010	798.62	0.000
17	0.371	−0.011	814.13	0.000
18	0.338	−0.012	827.17	0.000
19	0.306	−0.010	838.00	0.000
20	0.275	−0.010	846.89	0.000
21	0.245	−0.009	854.06	0.000
22	0.217	−0.007	859.75	0.000
23	0.190	−0.007	864.18	0.000
24	0.165	0.012	867.58	0.000

Figure 2: The ACF and PACF graph of the original time series.

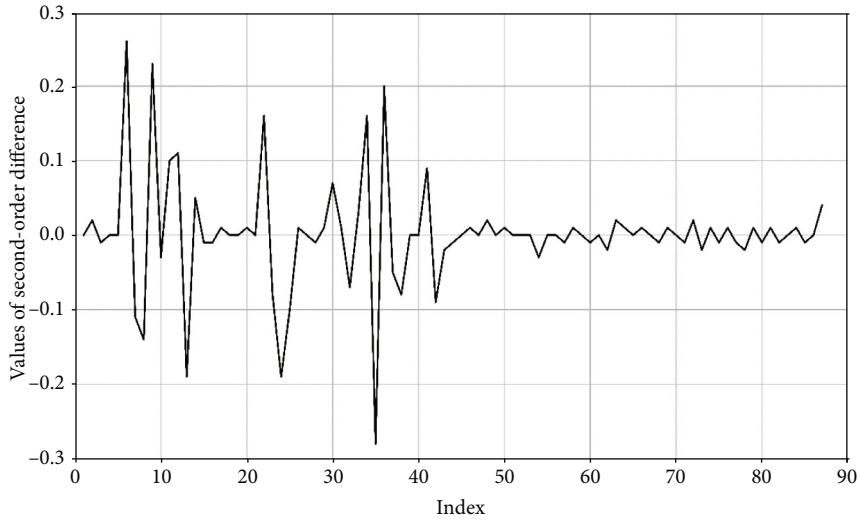
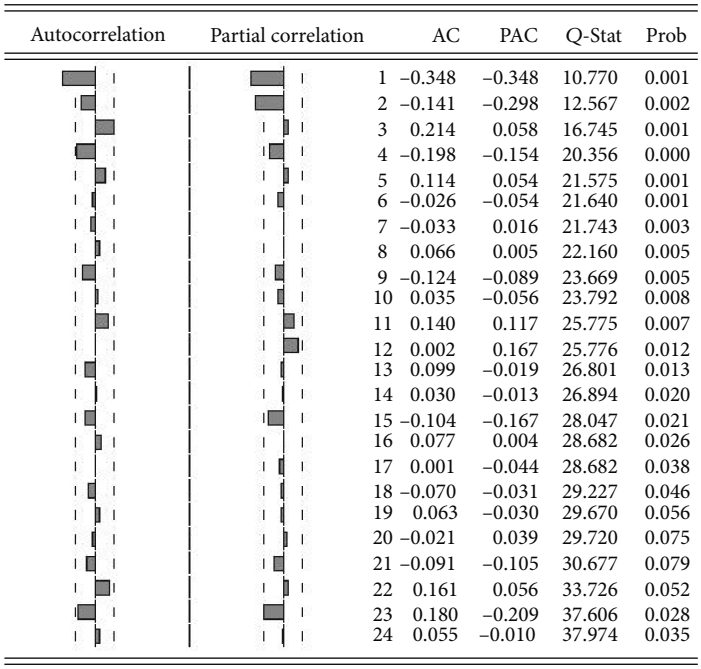


Figure 3: The curve of the second difference time series.

of differencing is an essential step to make time series stationary. The number of differencing required to make sequence stationary in this paper is gained by ndiffs() function of R programming language, and the term d of the ARIMA model is ultimately determined to be 2. Figure 3 is the second-difference graph of the original data; it can be seen intuitively that the curve fluctuates around 0 and the trend is stationary. The ACF and PACF graph of the second-difference of original data is displayed in Figure 4; as can be observed in the figure, the values of the function are sharply dropped down to 0 and fluctuated in the confidence interval. In addition, the measure of Augmented Dickey-Fuller test (ADF) is applied to identify the stationary or nonstationary processes of the second difference time series, the results are presented in Table 1, the significance level is set as 1%, 5%, and 10%, and the small p value suggests that the second difference time series is stationary.

4.2. ARIMA Model Construction. The autoregressive integrated moving average (ARIMA) model is based on the classical Box-Jenkins methodology for forecasting time series data. The precondition of establishing the ARIMA model is obtaining a stationary time series, and it has been fulfilled by the aforementioned differencing procedure. In this research, the terms of the ARIMA model are determined automatically by auto.arima function of R programming



Autocorrelation	Partial correlation		AC	PAC	Q-Stat	Prob
		1	−0.348	−0.348	10.770	0.001
		2	−0.141	−0.298	12.567	0.002
		3	0.214	0.058	16.745	0.001
		4	−0.198	−0.154	20.356	0.000
		5	0.114	0.054	21.575	0.001
		6	−0.026	−0.054	21.640	0.001
		7	−0.033	0.016	21.743	0.003
		8	0.066	0.005	22.160	0.005
		9	−0.124	−0.089	23.669	0.005
		10	0.035	−0.056	23.792	0.008
		11	0.140	0.117	25.775	0.007
		12	0.002	0.167	25.776	0.012
		13	0.099	−0.019	26.801	0.013
		14	0.030	−0.013	26.894	0.020
		15	−0.104	−0.167	28.047	0.021
		16	0.077	0.004	28.682	0.026
		17	0.001	−0.044	28.682	0.038
		18	−0.070	−0.031	29.227	0.046
		19	0.063	−0.030	29.670	0.056
		20	−0.021	0.039	29.720	0.075
		21	−0.091	−0.105	30.677	0.079
		22	0.161	0.056	33.726	0.052
		23	0.180	−0.209	37.606	0.028
		24	0.055	−0.010	37.974	0.035

Figure 4: The ACF and PACF graph of the second difference time series.

Table 1: The statistical results of the ADF test for the second difference time series.

		t-statistic	*p*-value
Augmented Dickey-Fuller test statistic		-10.05383	≤0.001
Test critical values:	1% level	-3.510259	
	5% level	-2.896346	
	10% level	-2.585396	

language, and the ARIMA model which with the minimum AIC (Akaike Information Criterion) score will be selected for future forecasts. In this context, the final model is determined as ARIMA(0, 2, 1) with a minimum AIC score -2.499856. The next step is to validate the applicability of the model using Box-Ljung statistics, if the calculated residuals are white noise, then the model will be determined as the definitive model for further forecasting; otherwise, a more suitable model needs to be found. The Box-Ljung test for the calculated residuals of the model is achieved using the Box.test function of R programming language; the statistical p value of the Box-Ljung test is 0.8993, and it indicates that this ARIMA model is suitable for further forecasting due to the p value of the Box-Ljung test being greater than 0.05. The final ARIMA model is utilized to conduct the further forecasting, and the prediction results are shown in Figure 5; the dashed line represents the prediction value; it can be seen that the fluctuation of the electricity consumption is on the rise in the next 28 days; the specific predicted values of the ARIMA model are presented in Table 2.

4.3. The Construction of Hybrid Model. It is a popular trend to combine the statistical model with machine learning methods to establish the electricity consumption prediction method. In the hybrid model, incorporating the residual error values into the predicted series deduces the better precision of the predictions [12]. The general construction process of the hybrid models is depicted in Figure 6. As presented in the following pipeline, the construction procedure of the hybrid model generally consists of two main steps; firstly, the ARIMA model is applied in this hybrid model as a well-known linear model; secondly, the residual error prediction is also considered in the hybrid model for the purpose of extracting the nonlinear features of the input data. The residual error is calculated from the ARIMA predictions and furthermore applied to the nonlinear model as input data. What is more, the procedures of normalization and renormalization are also contained in the design process with the aim of removing the influence of the magnitude levels. The obtained corrected predictions are taken as the ultimate results of the hybrid model and applied as the final predicted value.

In general, it is more effective to combine individual models for forecasting energy consumption. The hybrid method which combines with machine learning and univariate ARIMA method has been used more frequently because the hybrid method could benefit from both of them. Just as its name implies, the ARIMA-SVR algorithm is a combination of the ARIMA model and SVR model. Similarly, the ARIMA-GBR model was also established in the same way, and the performance comparison of the five models is presented in Table 2.

Figure 7 has depicted the forecasting results of the proposed model and relevant models with different predicted days. Within a prediction interval of 16 days, the output of the ARIMA-SVR model more closely approximates the

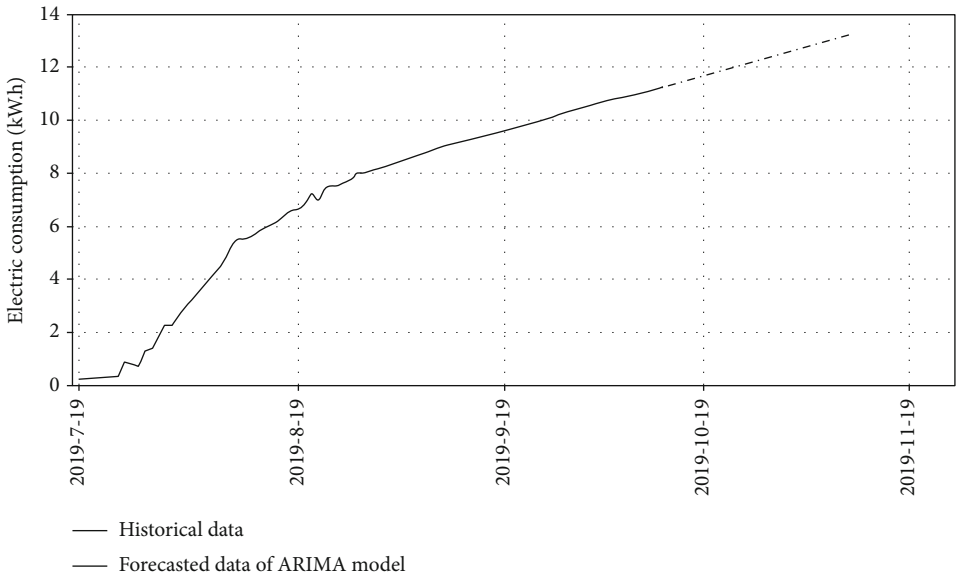


Figure 5: The prediction curve of the ARIMA model (dash line represents the prediction value).

Table 2: Forecasting results of the proposed model and relevant models.

Date	Actual values (kw/h)	Predicted values (kw/h)				
		ARIMA model	ARIMA-SVR model	ARIMA-GBR model	LSTM model	GRU model
2019-10-16	11.31	11.2915	11.3282	11.2899	11.2491	11.2279
2019-10-17	11.39	11.3629	11.4004	11.3614	11.3041	11.3220
2019-10-18	11.48	11.4344	11.4726	11.3554	11.5189	11.4989
2019-10-19	11.56	11.5059	11.5448	11.4269	11.4577	11.4398
2019-10-20	11.65	11.5773	11.617	11.4983	11.6105	11.5731
2019-10-21	11.71	11.6488	11.6892	11.6213	11.6519	11.6113
2019-10-22	11.77	11.7203	11.7614	11.7061	11.5522	11.601
2019-10-23	11.85	11.7917	11.8336	11.7775	11.6989	11.7306
2019-10-24	11.90	11.8632	11.9058	11.8490	11.8505	11.7955
2019-10-25	11.96	11.9347	11.9779	11.9205	11.8838	11.8988
2019-10-26	12.04	12.0061	12.0501	11.9919	11.8921	11.9203
2019-10-27	12.12	12.0776	12.1223	12.0634	11.8078	11.8065
2019-10-28	12.19	12.1491	12.1945	12.1349	11.9552	11.8758
2019-10-29	12.27	12.2205	12.2667	12.2446	11.8943	11.9488
2019-10-30	12.35	12.2920	12.3389	12.3160	12.041	12.0671
2019-10-31	12.40	12.3635	12.4111	12.3852	11.9898	12.0154
2019-11-1	12.44	12.4349	12.4833	12.4567	12.2121	12.6195
2019-11-2	12.48	12.5064	12.5555	12.5095	12.1576	12.5813
2019-11-3	12.53	12.5779	12.6276	12.5809	12.3366	12.553
2019-11-4	12.57	12.6494	12.6998	12.6524	12.4153	12.6426
2019-11-5	12.62	12.7208	12.7720	12.7239	12.5056	12.8841
2019-11-6	12.66	12.7923	12.8442	12.7953	12.6973	13.0932
2019-11-7	12.71	12.8638	12.9164	12.8668	12.6343	12.8843
2019-11-8	12.80	12.9352	12.9886	12.9383	12.7165	13.0761
2019-11-9	12.85	13.0067	13.0608	13.0097	12.5879	12.4027
2019-11-10	12.89	13.0781	13.133	13.0812	12.6287	12.5881
2019-11-11	12.94	13.1496	13.2052	13.1527	12.7034	12.7740
2019-11-12	12.99	13.2211	13.2774	13.2241	12.9237	13.1884

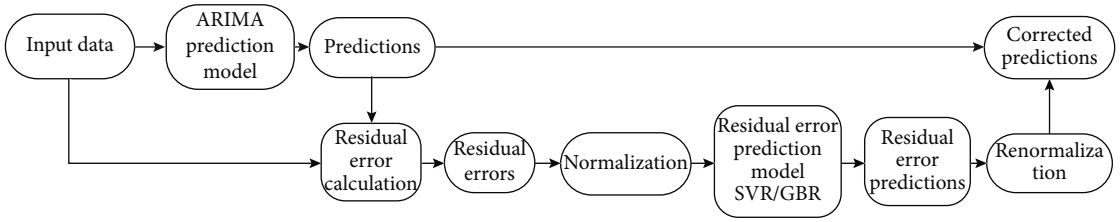


Figure 6: The flow chart of the hybrid model construction.

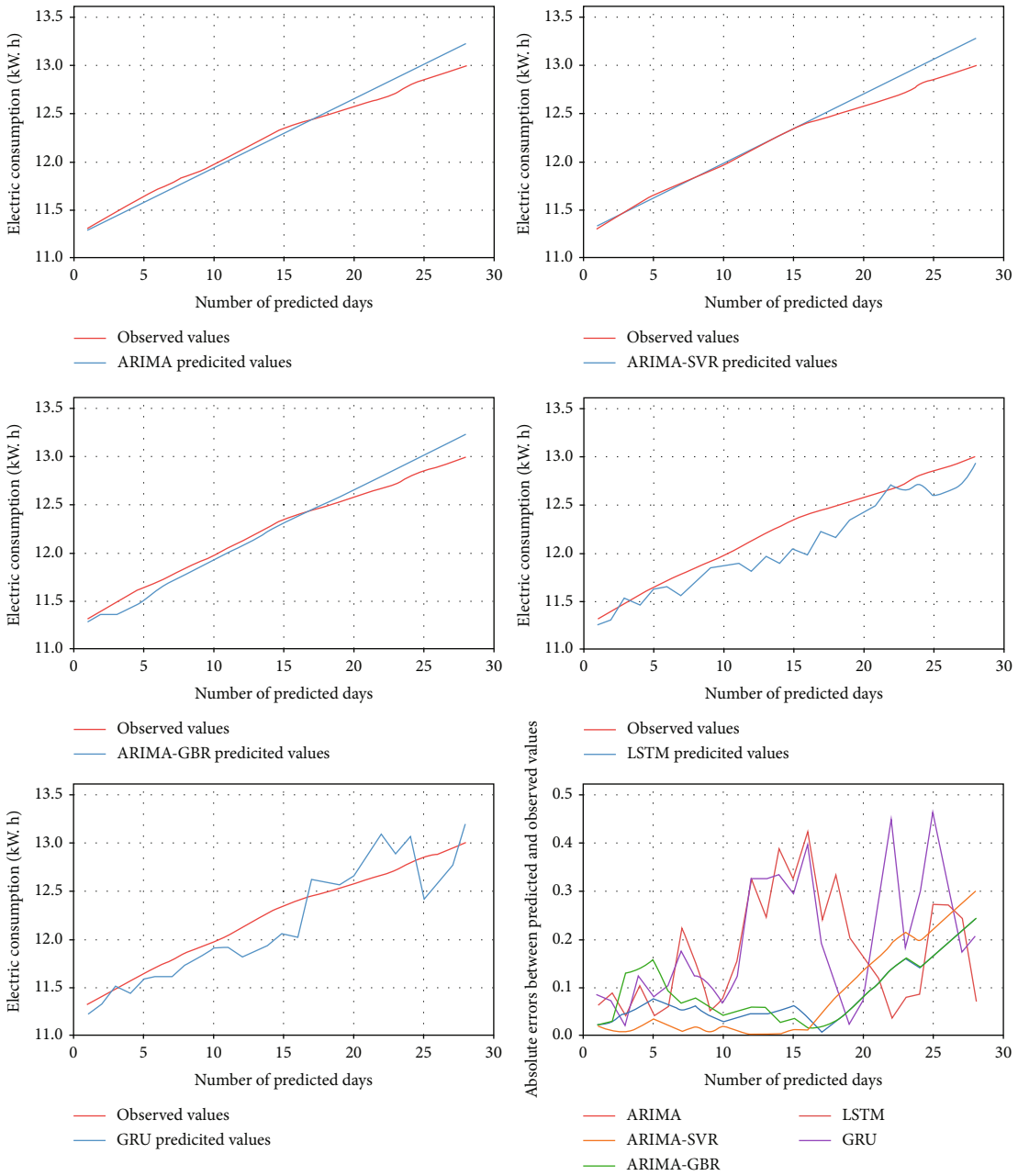


Figure 7: Forecasting results of the proposed model and relevant models.

measured electricity consumption than that other relevant models. The last subgraph above is the absolute error curve between the observed and predicted values; the yellow line represents the absolute error between the predicted values of the ARIMA-SVR model and the observed data; it fluctuates lower than the other curves stability until the predicted

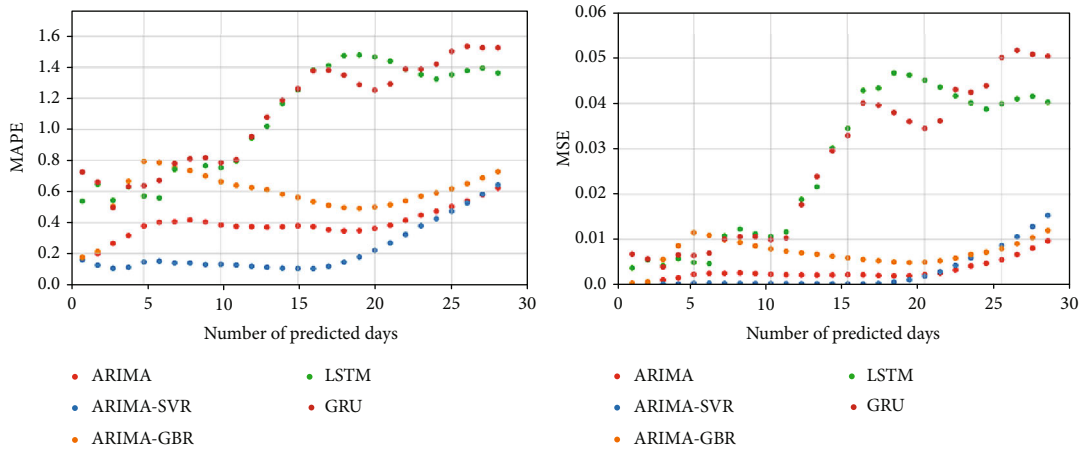


Figure 8: The MAPE and MSE of the five forecasting models.

point reached at 16 days; the absolute error curve of the single ARIMA model is under than that of ARIMA-SVR model when the predicted point beyond 16 days, and it almost overlapped with the error curve of the ARIMA-GBR hybrid model. The results come out to be that the hybrid model ARIMA-SVR could improve the prediction performance of the individual ARIMA model to a certain degree, and the drawback of the single linear model could be overcome when it is in conjunction with some nonlinear models which have the capability of capturing nonlinear features in the dataset. This is because the time series is often neither purely linear nor purely nonlinear. Thus using an individual model alone could not capture the data characteristics adequately. Besides, it is worth noting that the hybrid model could improve the accuracy of forecasting efficiently only in certain situations or in a certain prediction horizon. The hybrid models are more and more used by many researchers due to their excellent predictive performance.

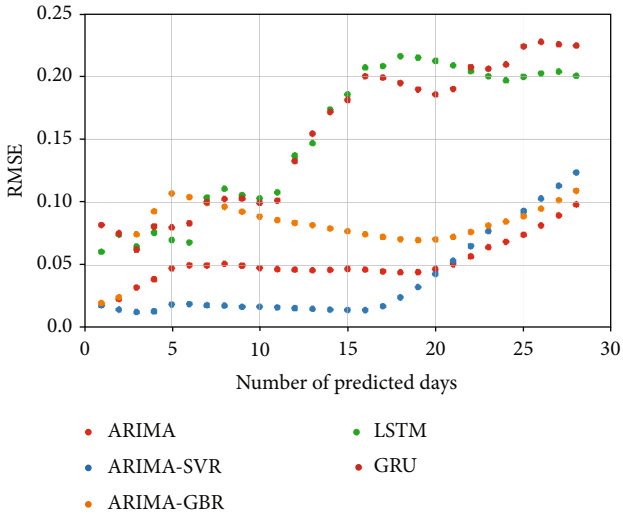


Figure 9: The RMSE of the five forecasting models.

5. Simulation Results and Discussion

The paper uses three evaluation criteria to assess the performance of the five models, respectively, are the mean squared error (MSE), the root mean square error (RMSE), and the mean absolute percentage error (MAPE); the formulations are detailed as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - y\wedge_i|^2, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - y_i\wedge)^2}, \tag{7}$$

$$\mathrm{MAPE} = \sum_{i=1}^{n} \left|\frac{y_i - \widehat{y}_i}{y_i}\right| \times \frac{100}{n}. \tag{8}$$

Here, y_i is the observed data, $\widehat{y}_i$ is the predicted value of the forecast model, and n is the number of the observed dataset.

Figures 8 and 9 have depicted the prediction errors MAPE, MSE, and RMSE of five forecasting models; it can be seen that the MAPE values of the ARIMA-SVR model are the lowest values among the other relevant models within the prediction horizon of 26 days, and the MSE and RMSE values of the ARIMA-SVR model are the lowest values within 20 predicted days. Also, the MSE and RMSE values of the ARIMA-SVR model maintain the lowest within 20 predicted days whereas the lowest region would be taken place by the errors of the single ARIMA model when prediction horizon beyond 20 days. Table 3 has presented the RMSE values of the five predictive models. The predictive performance of the LSTM and GRU models is not as good as the ARIMA-SVR model; the main reason is probably that the dataset is small in this paper. In addition, it is notable that the prediction performance of the ARIMA-SVR model is more stable compared to the other models, and the prediction results showed that the ARIMA-SVR model is a good choice when the electricity consumption is predicted within 20 days, and the ARIMA model is suitable for the prediction over 20 days.

Table 3: The root mean square error (RMSE) of the five predictive models.

Predicted days	Root mean square error (RMSE)				
	ARIMA model	ARIMA-SVR model	ARIMA-GBR model	LSTM model	GRU model
1	0.0185	0.0182	0.0201	0.0609	0.0821
2	0.0232	0.0148	0.0247	0.0745	0.0754
3	0.0324	0.0128	0.0747	0.0648	0.0625
4	0.039	0.0135	0.0928	0.0759	0.0809
5	0.0477	0.019	0.1072	0.0702	0.0801
6	0.0502	0.0194	0.1044	0.0683	0.0835
7	0.0501	0.0182	0.0996	0.1038	0.1003
8	0.0512	0.018	0.0966	0.1108	0.1029
9	0.0498	0.0171	0.0927	0.1058	0.103
10	0.0479	0.0172	0.0888	0.1032	0.0997
11	0.0468	0.0167	0.0859	0.108	0.1016
12	0.0465	0.016	0.0838	0.1372	0.1329
13	0.0461	0.0154	0.082	0.147	0.1546
14	0.0463	0.0148	0.0793	0.1736	0.1719
15	0.0472	0.0146	0.0771	0.1858	0.1814
16	0.0466	0.0144	0.0748	0.2071	0.2003
17	0.0452	0.0175	0.0726	0.2083	0.1991
18	0.0444	0.0246	0.0709	0.2163	0.195
19	0.0446	0.0328	0.07	0.2151	0.1898
20	0.0469	0.0432	0.0707	0.2125	0.1857
21	0.0508	0.0536	0.0726	0.2089	0.1902
22	0.057	0.0655	0.0766	0.2042	0.2075
23	0.0644	0.0772	0.0817	0.2004	0.2062
24	0.0688	0.0848	0.0848	0.1969	0.2096
25	0.0743	0.0932	0.0891	0.1999	0.224
26	0.0817	0.103	0.095	0.2026	0.2275
27	0.0898	0.1132	0.1018	0.204	0.2255
28	0.0983	0.1238	0.1094	0.2007	0.2246

The experiment is conducted on a PC with Intel Core i5-8300H CPU @2.30GHz, 8.00 GB RAM, and 64-bit operating systems. Exhibited graphs and hybrid models are implemented in Pycharm using Python language.

6. Conclusions

The results have indicated that the hybrid method ARIMA-SVR has great capability for forecasting the electricity consumption of the buildings; it is efficient for enhancing the accuracy of the electricity consumption prediction in certain conditions. In the prediction horizon of 20 days, the hybrid model ARIMA-SVR has significant superiority than the other four models while the ARIMA model is a better choice when the prediction horizon exceeds 20 days.

The commended hybrid ARIMA-SVR model in this paper has provided a new perspective of understanding a complex energy data structure. The general process of the hybrid model building is always decomposing the original time series into a stationary linear component and a fluctuant nonlinear residual. The limitations of the proposed model are that it highly dependents on time series and requires the univariate time series to be stationary or to be stationary after differencing; data preprocessing is an essential step before building the model. As we all know, the overall electricity consumption of buildings was also related to energy-using behaviors and climatic factors. In further works, we plan to consider some relevant eigenvalues such as temperature and humidity into model construction and focus on combining some popular algorithms testing on different size datasets; on the other hand, we will focus on extending the effective scope of the proposed model in future studies.

Forecasting electricity consumption in advance is of great importance in achieving energy conservation, and it could provide data support for policymakers and energy managers to make reasonable strategies to promote environmental protection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 52074305, in part by the National Natural Science Foundation of China under Grant 51874300 and 51874299, in part by the National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon under Grant U1510115, and in part by the Open Research Fund of Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, under Grant 20190902.

References

[1] Y. Sun, W. Hao, Y. Chen, and B. Liu, "Data-driven occupant-behavior analytics for residential buildings," *Energy*, vol. 206, p. 118100, 2020.

[2] E. Caro and J. Juan, "Short-term load forecasting for Spanish insular electric systems," *Energies*, vol. 13, no. 14, article 3645, 2020.

[3] S. L. Ho and M. Xie, "The use of ARIMA models for reliability forecasting and analysis," *Computers & Industrial Engineering*, vol. 35, no. 1-2, pp. 213–216, 1998.

[4] H. Tandon, P. Ranjan, T. Chakraborty, and V. Suhag, "Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future," 2020, https://arxiv.org/abs/2004.07859.

[5] A. Barman, "Time series analysis and forecasting of COVID-19 cases using LSTM and ARIMA models," 2020, https://arxiv.org/abs/2006.13852.

[6] R. Jamil, "Hydroelectricity consumption forecast for Pakistan using ARIMA modeling and supply-demand analysis for the year 2030," *Renewable Energy*, vol. 154, pp. 1–10, 2020.

[7] Y. Li, Y. Wei, and Z. Dong, "Will China achieve its ambitious goal?—forecasting the CO_2 emission intensity of china towards 2030," *Energies*, vol. 13, no. 11, article 2924, 2020.

[8] F. Kaytez, "A hybrid approach based on autoregressive integrated moving average and least- square support vector machine for long-term forecasting of net electricity consumption," *Energy*, vol. 197, article 117200, 2020.

[9] A. T. Eseye and M. Lehtonen, "Short-term forecasting of heat demand of buildings for efficient and optimal energy management based on integrated machine learning models," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7743–7755, 2020.

[10] T. Ahmad and H. Zhang, "Novel deep supervised ML models with feature selection approach for large- scale utilities and buildings short and medium-term load requirement forecasts," *Energy*, vol. 209, article 118477, 2020.

[11] C. Liu, B. Sun, C. Zhang, and F. Li, "A hybrid prediction model for residential electricity consumption using holt- winters and extreme learning machine," *Applied Energy*, vol. 275, article 115383, 2020.

[12] A. Tascikaraoglu and M. Uzunoglu, "A review of combined approaches for prediction of short-term wind speed and power," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 243–254, 2014.

[13] M. R. Kazemzadeh, A. Amjadian, and T. Amraee, "A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting," *Energy*, vol. 204, article 117948, 2020.

[14] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Japan Architectural Review*, vol. 3, no. 1, pp. 62–76, 2020.

[15] M. Ma and Z. Wang, "Prediction of the energy consumption variation trend in South Africa based on ARIMA, NGM and NGM-ARIMA Models," *Energies*, vol. 13, no. 1, p. 10, 2020.

[16] E. Gulay and O. Duru, "Hybrid modeling in the predictive analytics of energy systems and prices," *Applied Energy*, vol. 268, article 114985, 2020.

[17] K. Balachander and D. Paulraj, "ANN and fuzzy based household energy consumption prediction with high accuracy," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 1–15, 2020.

[18] T. Silveira Gontijo and M. Azevedo Costa, "Forecasting hierarchical time series in power generation," *Energies*, vol. 13, no. 14, article 3722, 2020.

[19] A. J. del Real, F. Dorado, and J. Durán, "Energy demand forecasting using deep learning: applications for the French grid," *Energies*, vol. 13, no. 9, article 2242, 2020.

[20] F. Prado, M. C. Minutolo, and W. Kristjanpoller, "Forecasting based on an ensemble autoregressive moving average - adaptive neuro - fuzzy inference system - neural network - genetic algorithm framework," *Energy*, vol. 197, article 117159, 2020.

[21] G. E. Box and G. C. Tiao, "Intervention analysis with applications to economic and environmental problems," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 70–79, 1975.

[22] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," *Advances in Neural Information Processing Systems*, vol. 9, pp. 281–287, 2008.

[23] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[24] M. Awad, R. Khanna, M. Awad, and R. Khanna, "Support vector regression," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pp. 67–80, Apress, Berkeley, CA, 2015.

[25] A. Chokor and M. El Asmar, "Data-driven approach to investigate the energy consumption of LEED-certified research buildings in climate zone 2B," *Journal of Energy Engineering*, vol. 143, no. 2, article 05016006, 2017.

[26] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.