Energy & Buildings 259 (2022) 111908

Contents lists available at ScienceDirect

Energy & Buildings

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

A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction

Irene Karijadi [a,b,*], Shuo-Yan Chou [a,c]

[a] Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan
[b] Department of Industrial Engineering, Widya Mandala Catholic University, Surabaya, Indonesia
[c] Taiwan Building Technology Center, National Taiwan University of Science and Technology, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 17 June 2021
Revised 17 January 2022
Accepted 22 January 2022
Available online 31 January 2022

Keywords:
Prediction
Decomposition
Time-series
RF
LSTM
Deep learning
Building

ABSTRACT

An accurate method for building energy consumption prediction is important for building energy management systems. However, building energy consumption data often exhibits nonlinear and nonstationary patterns, which makes prediction more difficult. This study proposes a hybrid method of Random Forest (RF) and Long Short-Term Memory (LSTM) based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to predict building energy consumption. In the first stage of our proposed method, the original energy consumption data is transformed into several components using CEEMDAN. Then, RF is used to predict the component with the highest frequency, and the remaining components are predicted using LSTM. In the last stage, the prediction results of all components are combined to obtain the final prediction results. The proposed method has been tested using real-world building energy consumption data. The experimental results demonstrate that the proposed method achieves better performance than the benchmark methods used for comparison.

1. Introduction

The building sector is one of the main energy consumers [1]. It accounts for 35% of global energy consumption and contributes 38% of total CO2 emissions [2]. Since 2000, energy consumption in the building sector has gradually increased with an annual growth rate of 1.1% [3], and it is predicted to continue rising over the next few decades [4,5]. The rise of building energy consumption has been driven by population growth and increased demands to build comfortable environments [6,7]. This increased use of energy consumption in the building sector raises concerns about supply issues and global environmental impacts [8]. Therefore, energy efficiency is needed in this sector to reduce carbon emissions and lessen the problems related to supply. Predicting energy consumption plays a prominent role in improving building energy efficiency. It serves as a basis for many advanced building energy management techniques, such as safety monitoring [9], demand response [10], and optimization control [11].

Based on the time scale of the prediction, energy prediction can be divided into four categories: long-term (a year or more), medium-term (between one week and one year), short-term (from one hour to one week), and very short-term prediction (from a few minutes to less than one hour) [12–14]. Long-term and medium-term predictions are critical for long-term and strategic planning [12]. They are often used as a reference to determine system capacity and system maintenance [15]. Short-term and very short-term predictions are beneficial for energy management. Many decisions related to energy management can be made based on short-term and very short-term predictions, such as peak load shaving, optimal energy scheduling, and demand-side management [16]. This study focuses on improving one-hour-ahead building energy prediction. The one-hour-ahead building energy prediction will provide an accurate baseline to estimate the impact of demand response measures on the buildings and for optimizing the local generator's schedule [12].

Over the years, various techniques have been developed to predict energy consumption, including statistical methods and machine learning methods. The statistical-based methods such as multiple linear regression [17], exponential smoothing [18], and Auto-Regressive Integrated Moving Average (ARIMA) [19] are able to fit the linear relationship in the data. However, these statistical methods are inadequate when dealing with nonlinear time series data. Machine learning-based techniques such as Random Forest (RF) [20,21]. Neural Network [22], Support Vector Regression

* Corresponding author at: Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan.
E-mail address: irenekarijadi92@gmail.com (I. Karijadi).

https://doi.org/10.1016/j.enbuild.2022.111908
0378-7788/

(SVR) [23–25], and Deep Learning with Long Short-Term Memory (LSTM) [26–28] have also been widely used in the field of building energy prediction due to their capability to model nonlinear series in the data.

Predicting energy consumption at an individual building level is quite challenging. It often exhibits nonstationary and high volatility patterns [28]. Thus, the single prediction methods mentioned earlier may be insufficient to capture all the patterns, limiting their effectiveness in achieving precise prediction. For this reason, several researchers have proposed hybrid methods that combine the decomposition technique with prediction methods to improve prediction accuracy. Their results have shown the effectiveness of the decomposition technique in improving prediction results [29–31]. For instance, Liu et al. [29] proposed a hybrid method based on EMD and SVR to predict energy consumption in an office building. Zheng et al. [30] combined the EMD and LSTM for short-term load prediction. Empirical mode decomposition (EMD) is a decomposition method that is effective in dealing with nonlinear and nonstationary time series data [32]. By adopting EMD to preprocess the original data, the performance of the prediction method can be effectively enhanced [29–31]. Despite its effectiveness, EMD suffers from mode mixing problems caused by intermittency signals [33]. To solve this issue, Wu and Huang introduced an enhancement version of EMD named Ensemble Empirical Mode Decomposition (EEMD) [34]. In EEMD, additional noise is added to solve the mode mixing problem. However, the decomposition results produced by EEMD would be contaminated by some residual noises, particularly when the number of ensemble trials is relatively low [35]. The noise can be eliminated by increasing the number of ensemble trials, but the calculation time will also increase [36]. Torres et al. [37] proposed the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm to improve the performance of EEMD. CEEMDAN adds adaptive noise in every stage of EMD, which can effectively improve decomposition results and reduce the computational time [37].

Although several studies have employed hybrid methods based on the decomposition method, these hybrid methods do not consider the unique characteristics of each component. They use identical prediction methods to predict all components. Hence, this study fills the gap by considering the characteristics of each component and combining different prediction methods to predict different components. In this study, a hybrid method of Random Forest (RF), Long Short-Term Memory (LSTM) based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is proposed to improve the accuracy of building energy prediction. The idea of the proposed method is to decompose the original data into several components. Then, we take into account the characteristics of each component and utilize different prediction methods to predict different components. By considering the different characteristics of each component, the corresponding prediction result can be further enhanced.

In the first stage of our approach, CEEMDAN is utilized to decompose the original data into several components. Compared with other decomposition methods, such as EMD and EEMD, CEEMDAN has shown better performance on the decomposition of the nonstationary series [38]. So, CEEMDAN is chosen as the decomposition approach in this study. After CEEMDAN decomposes original data into several components, different prediction methods are used to predict different components that correspond to their characteristics. The highest frequency component, which is the first component, is modeled and predicted using RF. The other components are predicted using LSTM. In the final stage, the prediction results of each component are integrated using summation to achieve the final prediction result. To the best of our knowledge, the application of hybrid RF-LSTM based on CEEMDAN for predicting building energy consumption has never been investigated before.

2. Theoretical background

2.1. Complete Ensemble Empirical Mode decomposition with Adaptive Noise (CEEMDAN)

CEEMDAN decomposes nonlinear and nonstationary series into several relatively stationary components. The components will be named as Intrinsic Mode Functions IMF_k. We define$E_j(\cdot)$ as the function to produce the j - th mode obtained by EMD procedure and ε_i is the amplitude coefficients of the white noise series. The decomposition process of CEEMDAN algorithm can be seen in Table 1.

2.2. Random Forest (RF)

Random Forest (RF) is a combination of many decision trees that can be used to solve classification and regression problems [39]. RF as an ensemble learning method works differently compared with other machine learning methods, such as ANN or SVR. ANN or SVR builds a global model from the original data, while RF builds several models and combines their results. This may achieve better accuracy, especially when handling complex systems [40]. The structure of RF for regression can be seen in Fig. 1.

The detailed process of the RF algorithm is as follows [41]:

Step 1: Create n bootstrap sample sets from the original dataset
Step 2: In every bootstrap sample set, generate an unpruned regression tree with the following modification: For each node, randomly sample p features from all the input features. After that, select the best split from p features, where p is less than the number of all input features (m).
Step 3: Obtain the random forest new output prediction by averaging the outputs of n regression trees when new input is fed into the model.

In the RF algorithm, there are two predetermined parameters: the number of trees n and the number of features for each node p. Generally, RF is not sensitive to the selection of these two parameters [40,42].

Recently, a new model based on RF called Deep Forest was proposed by Zhou and Feng [43] to solve the classification problem. Deep forest is a multilayer structure with each layer consisting of many random forests. The deep forest model has been shown to outperform deep neural networks to some extent [44]. However, the current deep forest is inefficient, lacks scalability[45], and the model may encounter an overfitting issue [44]. Therefore, in this research, we decided not to adopt deep forest and use random forest instead, as it has been proven to be a robust and efficient algorithm for regression tasks [46].

2.3. Long Short-Term Memory (LSTM)

The energy consumption data can be regarded as time-series data, consisting of a series of observations recorded sequentially over time [47]. Recurrent Neural Network (RNN) is an enhanced version of a traditional neural network, and it is primarily designed to predict sequential data [48]. RNN uses a recurrent cell whose activation at each time is dependent on the activation at the earlier time to handle sequential data [49]. However, RNN has limitations in capturing long-term dependencies in the sequential data. To mitigate the problems of learning long-term dependencies, Long

Table 1
CEEMDAN algorithm [37].

Algorithm 1 CEEMDAN Algorithm

Step 1. White noise series w^i with $\mathcal{N}(0,1)$ is added to the original data $x[n]$:
$x[n]+\varepsilon_0 w^i[n]$ (1)
Step 2. Decompose by EMD I realizations $x[n]+\varepsilon_0 w^i[n]$ using EMD shifting procedures [32] to obtain their first modes and compute:
$\widetilde{IMF}_1[n]=\frac{1}{I}\sum_{i=1}^{I} IMF_1^i[n]=\overline{IMF}_1[n]$ (2)
Step 3. Calculate the first stage residual:
$r_1[n]=x[n]-\widetilde{IMF}_1[n]$ (3)
Step 4. Decompose $r_1[n]+\varepsilon_1 E_1(w^i[n]), i=1,\cdots,I$ until their first EMD mode is obtained. Then the second mode $\widetilde{IMF}_2[n]$ would be computed as
$\widetilde{IMF}_2[n]=\frac{1}{I}\sum_{i=1}^{I} E_1(r_1[n]+\varepsilon_1 E_1(w^i[n]))$ (4)
Step 5. For $k=2,\ldots K$, calculate the k - th residue:
$r_k[n]=r_{(k-1)}[n]-\widetilde{IMF}_k[n]$ (5)
Step 6. Decompose $r_k[n]+\varepsilon_k E_k(w^i[n]), i=1,\cdots,I$ until their first EMD mode is obtained and the $(k+1)$-th mode can be computed as
$\widetilde{IMF}_{(k+1)}[n]=\frac{1}{I}\sum_{i=1}^{I} E_1(r_k[n]+\varepsilon_k E_k(w^i[n]))$ (6)
Step 7. For the next k perform step 5.
Step 5 to Step 7 will be repeated to obtain the IMF components until the residual is a monotony function and cannot be decomposed by EMD. The final decomposition results of original data can be expressed as
$x[n]=\sum_{k=1}^{K}\widetilde{IMF}_k+R[n]$ (7)
with K as the total number of modes.

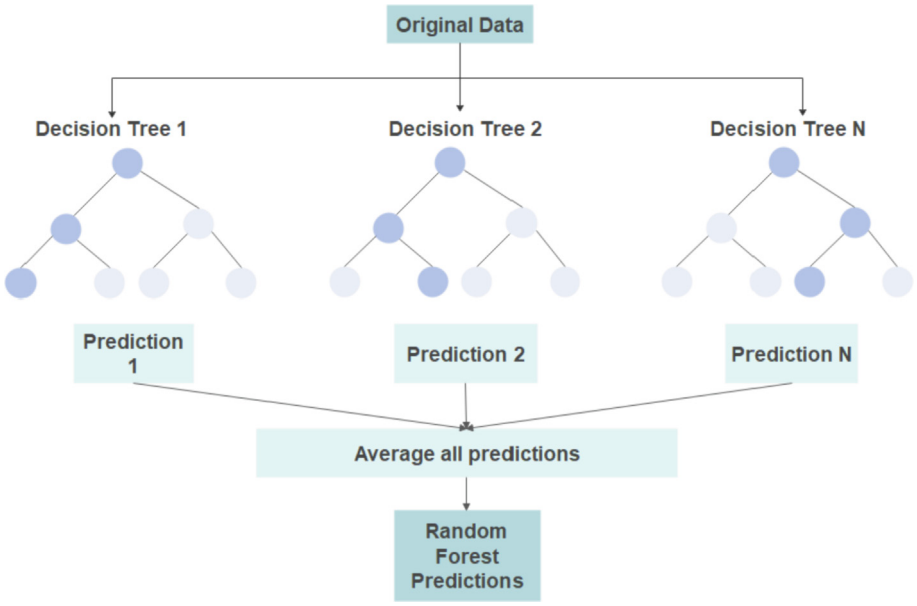

Fig. 1. Structure of Random Forest for a regression problem.

Short-Term Memory (LSTM), an improved version of RNN, was introduced by Hochreiter & Schmidhuber in 1997 [50]. LSTM improves the standard recurrent cell memory capability by introducing the "gates" mechanism into the cell [51]. These gates in the LSTM cell cooperate with each other to regulate how much information should be kept and how much should be forgotten. These structures allow the network to neglect less useful historical information and retain important information over longer periods. Hence, LSTM is capable of capturing long-term dependency in sequential data [52], and it has been widely used to deal with classification and regression problems of time series data. The structure of LSTM can be visualized as shown in Fig. 2 [53].

Specifically, the LSTM cell structure consists of three gates: the forget gate, the input gate, and the output gate [52]. These three gates control the information flow. The computational process in LSTM starts with the decision on what information should be preserved or removed in the forget gate, which can be expressed as

$$f_t=\sigma(W_f.[h_{t-1},x_t]+b_f) \tag{8}$$

The value comes out between 0 and 1, where a value closer to 1 means keeping the information and a value closer to 0 means forgetting the information. The input layer will decide which new information will be added to the cell state. It consists of two procedures. First, the previous hidden state and current input are passed into a sigmoid function:

$$i_t=\sigma(W_i.[h_{t-1},x_t]+b_i) \tag{9}$$

Then, the hidden state and current input are also passed into the tanh function to control how much information is added:

$$C_t=tanh(W_C.[h_{t-1},x_t]+b_C) \tag{10}$$

The next step is updating the cell state of the memory cells. This can be done by multiplying the old state C_{t-1} by f_t and add $i_t*\widehat{C_t}$ which can be expressed as

$$C_t=f_t*C_{t-1}+i_t*C_t \tag{11}$$

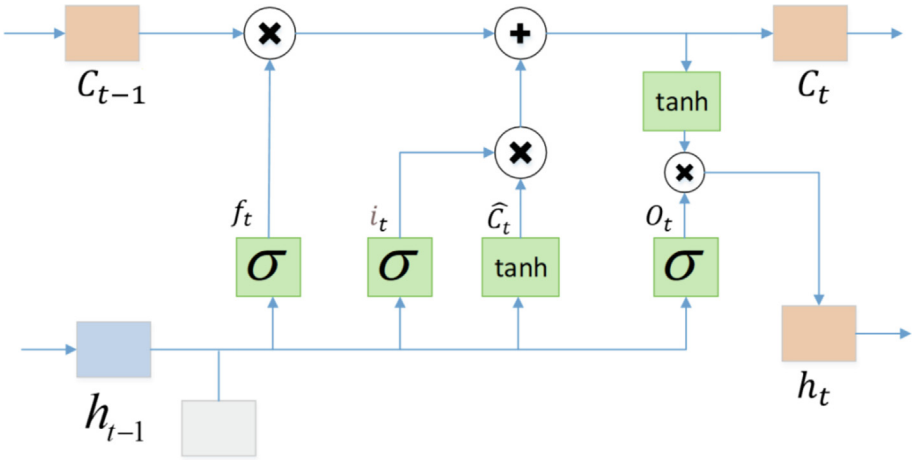

Fig. 2. LSTM structure.

Then the output is obtained by multiplying the output and the new cell state as defined in Eq. (12), and the new hidden state is computed as

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{12}$$

$$h_t = o_t * tanh(C_t) \tag{13}$$

whereσ is the sigmoid function$\sigma = \frac{1}{1+e^{-z}}$, tanh is the tanh function $\frac{e^x - e^{-x}}{e^x + e^{-x}}$. W_f,W_t, W_c,W_σ are the weight metrics, and b_f, b_i,b_c, b_o are the bias vectors.

Besides LSTM, RNN also has another variant named Gated Recurrent Unit (GRU) [54]. GRU is less complex than LSTM because the GRU cell integrates the forget gate and input gate as an update gate [51]. Thus, the GRU cell has only two gates: an update gate and a reset gate. However, GRU has a slow convergence rate [55], and it is less powerful than the original LSTM [51]. For this reason, this study will further focus on the recurrent network architectures based on LSTM cells.

3. Framework of the proposed method

In this study, a hybrid method that consists of Random Forest (RF), Long Short-Term Memory (LSTM), and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) for building energy consumption prediction is introduced. In the first stage, the original data is decomposed using CEEMDAN into several Intrinsic Mode Function (IMF) components. Each IMF component generated from CEEMDAN has different characteristics and is arranged from the highest to the lowest frequency. The first component (IMF1) represents the highest frequency component. The remaining frequency IMFs reflect the periodic patterns or seasonality of the data. The last IMF component, also known as the residual, is the lowest frequency component. It also represents the overall trend of the data.

The next stage is data prediction, where different prediction methods are used to predict different components corresponding to their characteristics. In order to predict the first IMF component, which is the most complex and highly fluctuating series, the RF prediction method is selected due to its ability to predict complex time series [40,56] and its robustness to outliers and noise [57]. To predict the remaining IMFs, which can be seen as the periodic component of the original data, LSTM is adopted. LSTM is chosen because LSTM has performed well in predicting periodic patterns in time series data [58,59]. Therefore, it is a good choice for predicting the remaining IMFs. For the residual component, which represents the long-term trend of the data, LSTM is adopted to predict the residual as LSTM is able to learn the trend of the time series [60]. In the last stage, the prediction outputs of all components are aggregated using summation to obtain final prediction results. The block diagram of the proposed method is shown in Fig. 3.

4. Experimental results

4.1. Data

In this study, we used a public dataset from the Building Data Genome project [61]. The Building Data Genome project has collected data from whole-building electrical meters, which includes the heating system in the building [61]. Hourly energy consumption data from March to May 2015 of five different buildings were analyzed in this study. The statistical information of these five buildings is presented in Table 2. Fig. 4 represents energy consumption profiles for each building at an hourly resolution. From Fig. 4, it can be observed that each building has different energy consumption patterns, which exhibit random and nonlinear patterns.

4.2. Evaluation metrics

Mean Absolute Percentage Error (MAPE), Root Means Square Error (RMSE), and Mean Absolute Error (MAE) are used as our evaluation metrics. The formula for MAPE, RMSE, and MAE are

$$MAPE\% = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t' - y_t}{y_t}\right| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t' - y_t)^2} \tag{15}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t' - y_t| \tag{16}$$

4.3. Experimental settings

In this study, a hybrid RF-LSTM method based on CEEMDAN is proposed to predict energy consumption in the building. CEEMDAN is used in this study to decompose original data into a number of Intrinsic Mode Functions (IMF) and a residual. This study used the pyEMD package [62] for implementing the CEEMDAN. We used

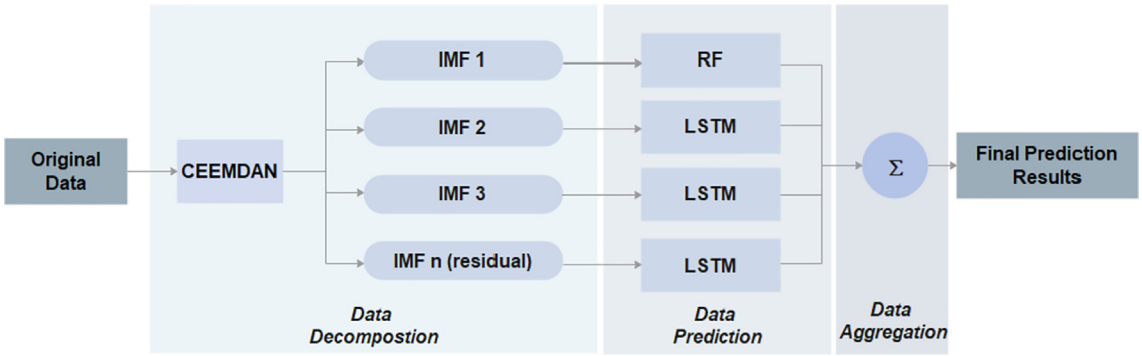

Fig. 3. Block Diagram of the Proposed Method.

Table 2
Statistical information.

Building Name	Building type	Gross Floor Area (ft^2)	Minimum value (kW)	Maximum value (kW)	Mean value (kW)	Standard deviation (kW)
Prince	University Dormitory	87,661	19.68	58.90	39.26	6.71
Christy	University Laboratory	28,084	23.78	64.38	41.36	7.71
Abby	University Classroom	9309	15.53	49.41	34.11	7.57
Abigail	Office	9309	3.39	19.11	6.89	3.13
Jaden	Primary/Secondary School Classroom	9703	5.87	24.03	5.87	5.46

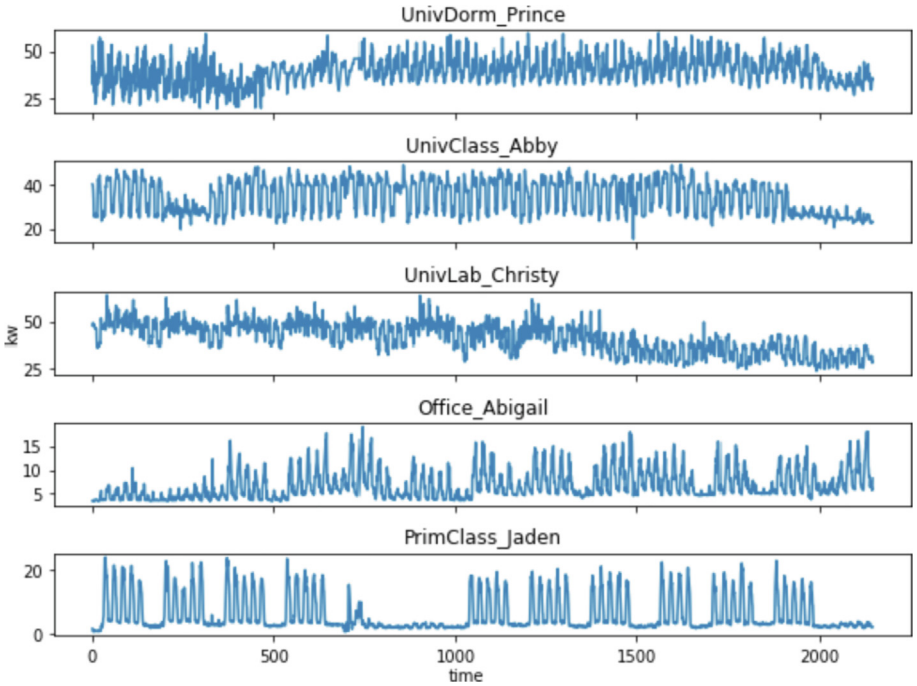

Fig. 4. Hourly energy consumption data of five different buildings.

scikit-learn to build RF prediction method [63] and Keras [64] in Python to implement the LSTM. We set the number of trees in RF to 100, which is the default value in the scikit-learn package [64], and the feature number of each node is set to 8, as it is suggested to be one-third of the feature's number [65].

For LSTM, this study used Adam optimizer with its recommended learning rate value of 0.001 [66]. Adam was chosen to optimize the model as it is computationally efficient and has prominent performance compared with other stochastic optimization methods [67,68]. Based on the literature, we trained our method for 100 epochs [69] and chose 64 as the batch size because the common value is the power of 2 [70]. Different combinations of LSTM hidden neurons with candidate values of 16,32 and 64 were tested. We found that the LSTM network model with 64 hidden neurons outperforms other configurations. Therefore, the architecture used in this study is a one-layer LSTM with 64 hidden neurons. This study conducts a one-hour ahead prediction (X_t), whose input includes the previous 24-hour energy consumption (X_{t-1} to X_{t-24}). We divided the data into training and test data sets, where 80% of data was used as training data sets, and 20% of the data was used as

test data sets. The experiments were performed on Intel Core i3-8130U CPU, 2.20 GHz, with a memory size of 4.00 GB.

5. Results

In the first stage of our proposed method, the original hourly energy consumption data was first decomposed using CEEMDAN. Fig. 5 shows the decomposition results for one of the buildings. As seen in Fig. 5, the frequency of each IMF obtained from the CEEMDAN process is arranged from the highest frequency to the lowest frequency. The first IMF component shows a highly irregular pattern; IMF 2 to IMF 7 show a periodic and more regular pattern. The last IMF component (IMF 8) shows the general trend of the data.

After CEEMDAN decomposes original data, the first IMF is predicted using RF, and the other IMFs are predicted using LSTM. In the final stage, the prediction results of each component are aggregated using summation to obtain the final prediction result. Fig. 6 visualizes the result obtained from the proposed method. As can be observed from Fig. 6, the predicted lines of the proposed method are close to the actual values line with small deviations, which means the proposed method can predict accurately.

The performance of the proposed hybrid RF-LSTM based on CEEMDAN for building energy prediction was compared with other prediction methods, including Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network, Long Short-Term Memory (LSTM), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-Random Forest (CEEMDAN-RF), and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise - Long Short-Term Memory (CEEMDAN-LSTM). According to the results in Table 3, the proposed method has the lowest error and has the best prediction accuracy among the benchmark methods.

To further measure the improvement of the proposed method in comparison to other benchmarking methods, three improvement percentage metrics are used in this study [71]. Percentage improvement of MAPE, RMSE, and MAE between two methods can be respectively calculated by:

$$P_{\mathrm{MAPE}} = \left| \frac{MAPE_1 - MAPE_2}{MAPE_1} \right| \tag{17}$$

$$P_{\mathrm{RMSE}} = \left| \frac{RMSE_1 - RMSE_2}{RMSE_1} \right| \tag{18}$$

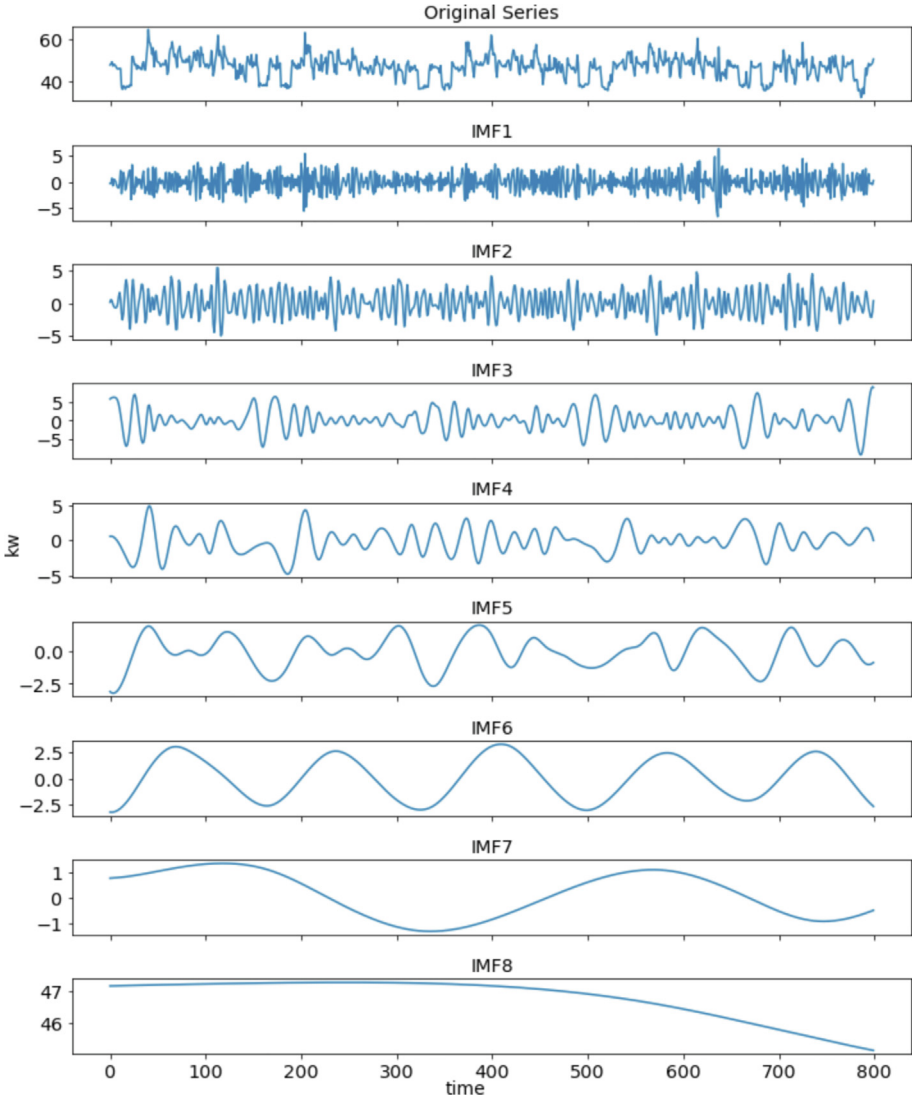

Fig. 5. Decomposition results for University Laboratory building.

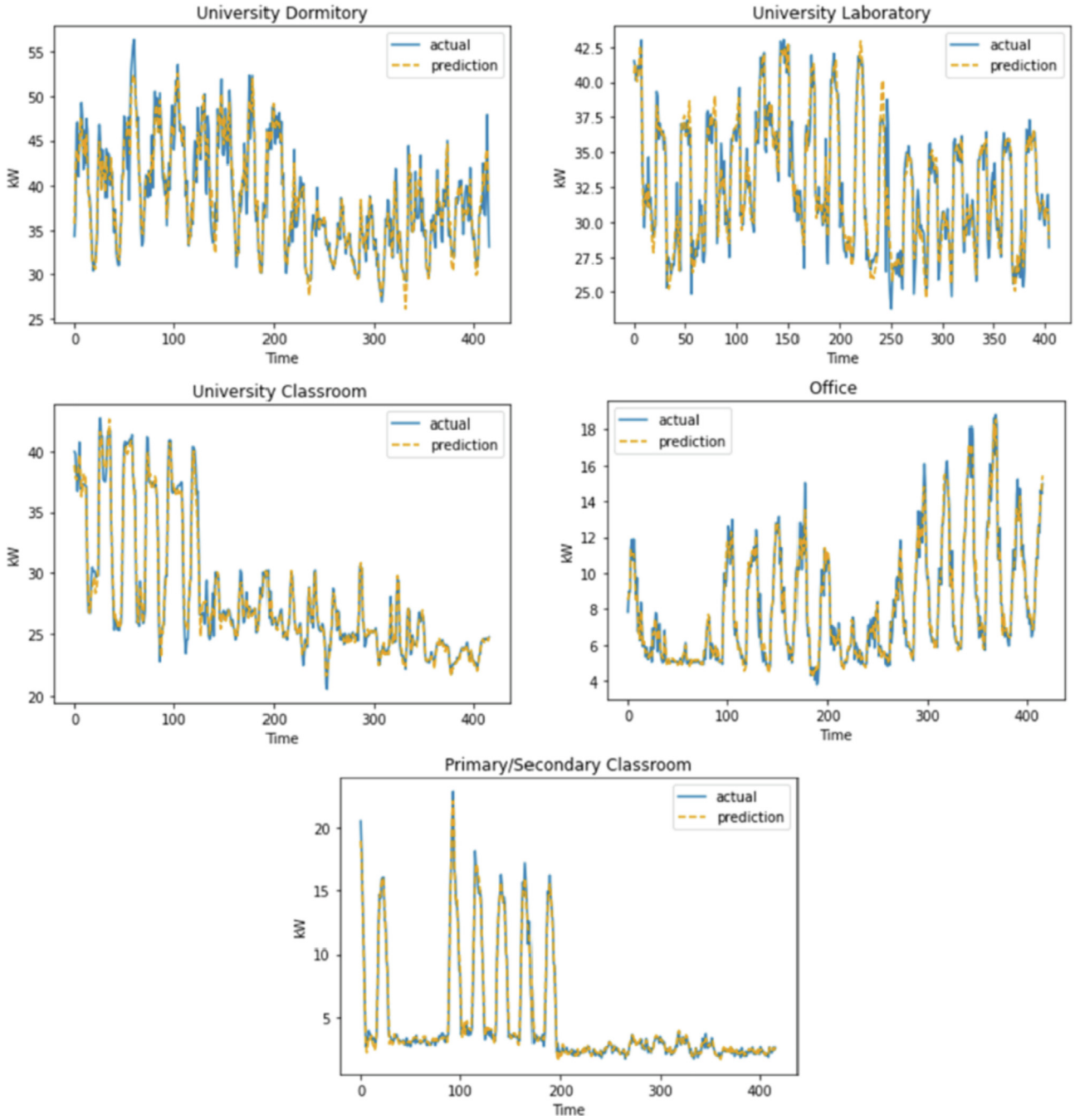

Fig. 6. Prediction results using proposed hybrid RF-LSTM based CEEMDAN method.

$$P_{\mathrm{MAE}} = \left| \frac{MAE_1 - MAE_2}{MAE_1} \right| \tag{19}$$

The percentages of error improvement compared with other benchmarking methods are summarized in Table 4.

Based on performance results listed in Table 3 and Table 4, we can observe that:

1. Hybrid-based CEEMDAN methods (CEEMDAN-RF, CEEMDAN-LSTM, and the proposed CEEMDAN-RF-LSTM method) perform better than other single prediction methods (such as LR, SVR, ANN, RF, and LSTM). This indicates that by implementing CEEMDAN to decompose original data, the prediction accuracy can be significantly enhanced. CEEMDAN is suitable for processing and reducing the nonstationary pattern that existed in the original building energy consumption data. By decomposing nonstationary original energy consumption data into several relatively stationary components, the prediction accuracy can be improved.
2. Compared with the other two hybrid CEEMDAN methods (CEEMDAN-RF and CEEMDAN-LSTM), which use identical prediction methods to predict all components, the proposed method performs better. This shows the effectiveness of combining different prediction methods in predicting IMF components. By considering the different characteristics of each component and using different methods to predict each component, the corresponding prediction result can be further enhanced.

Table 3
Performance results of different prediction methods.

Building	Evaluation Metrics	Prediction Methods							
		LR	SVR	ANN	RF	LSTM	CEEMDAN RF	CEEMDAN LSTM	Proposed Method (CEEMDAN-RF-LSTM)
University Dormitory	MAPE (%)	6.097	6.487	6.719	6.145	6.829	3.955	4.081	**3.511**
	RMSE	3.091	3.386	3.389	3.172	3.468	2.008	2.070	**1.761**
	MAE	2.401	2.590	2.641	2.436	2.678	1.551	1.587	**1.369**
	Running Time (s)	*0.336*	*0.380*	*1.944*	*1.118*	*11.987*	*37.393*	*105.333*	*103.585*
University Laboratory	MAPE (%)	5.859	6.419	6.135	6.273	6.385	3.566	3.509	**3.191**
	RMSE	2.580	2.707	2.659	2.715	2.627	1.466	1.429	**1.293**
	MAE	1.865	2.093	2.005	2.046	2.063	1.132	1.116	**1.014**
	Running Time (s)	*0.236*	*0.376*	*1.534*	*1.120*	*6.681*	*31.269*	*86.682*	*75.807*
University Classroom	MAPE (%)	3.656	5.807	5.983	6.873	5.169	3.084	2.123	**1.965**
	RMSE	1.629	2.529	2.402	2.675	2.169	1.332	0.866	**0.815**
	MAE	1.062	1.780	1.748	2.045	1.530	0.920	0.613	**0.570**
	Running Time (s)	*0.536*	*0.352*	*1.943*	*1.160*	*9.038*	*32.142*	*85.869*	*81.056*
Office	MAPE (%)	11.169	9.307	11.291	10.251	9.792	6.165	6.354	**5.331**
	RMSE	1.218	1.106	1.148	1.084	1.124	0.650	0.669	**0.570**
	MAE	0.902	0.770	0.876	0.805	0.806	0.491	0.499	**0.430**
	Running Time (s)	*0.277*	*0.290*	*1.638*	*1.162*	*6.344*	*32.364*	*83.035*	*81.803*
Primary/ Secondary Classroom	MAPE (%)	15.508	14.671	14.464	18.396	16.652	10.944	8.550	**7.164**
	RMSE	1.040	1.150	0.902	1.342	0.957	0.750	0.524	**0.467**
	MAE	0.651	0.647	0.578	0.785	0.622	0.455	0.342	**0.299**
	Running Time (s)	*0.319*	*0.174*	*0.829*	*1.215*	*7.894*	*42.319*	*98.648*	*95.249*

*bold values represent the best result for each metric on each building.

Table 4
Improvement percentage of Hybrid RF-LSTM based on CEEMDAN compared with other benchmarking methods.

Building	Improvement Percentage Metrics	Proposed Method vs. LR	Proposed Method vs. SVR	Proposed Method vs. ANN	Proposed Method vs. RF	Proposed Method vs. LSTM	Proposed Method vs. CEEMDAN RF	Proposed Method vs. CEEMDAN LSTM
University Dormitory	$P_{MAPE}(\%)$	42.4%	45.87%	47.24%	42.86%	48.58%	11.22%	13.95%
	$P_{RMSE}(\%)$	43.03%	47.99%	48.04%	44.48%	49.22%	12.28%	14.91%
	$P_{MAE}(\%)$	42.97%	47.13%	48.14%	43.79%	48.86%	11.7%	13.%
University Laboratory	$P_{MAPE}(\%)$	45.54%	50.3%	47.99%	49.14%	50.03%	10.54%	9.08%
	$P_{RMSE}(\%)$	49.86%	52.22%	51.35%	52.37%	50.77%	11.78%	9.49%
	$P_{MAE}(\%)$	45.6%	51.54%	49.4%	50.41%	50.84%	10.37%	9.09%
University Classroom	$P_{MAPE}(\%)$	46.25%	66.17%	67.16%	71.01%	61.99%	36.29%	7.46%
	$P_{RMSE}(\%)$	49.93%	67.76%	66.05%	69.22%	62.41%	38.79%	5.84%
	$P_{MAE}(\%)$	46.32%	67.98%	67.40%	71.75%	62.76%	38.03%	7.06%
Office	$P_{MAPE}(\%)$	52.27%	42.72%	52.79%	48.00%	45.56%	13.53%	16.10%
	$P_{RMSE}(\%)$	53.17%	48.44%	50.34%	47.38%	49.24%	12.21%	14.70%
	$P_{MAE}(\%)$	52.27%	44.08%	50.89%	46.64%	46.58%	12.40%	13.79%
Primary/ Secondary Classroom	$P_{MAPE}(\%)$	53.81%	51.17%	50.47%	61.06%	56.98%	34.54%	16.21%
	$P_{RMSE}(\%)$	55.14%	59.43%	48.26%	65.23%	51.24%	37.75%	10.89%
	$P_{MAE}(\%)$	54.07%	53.81%	48.26%	61.96%	52.01%	34.40%	12.62%

3. The running time of the hybrid-based CEEMDAN methods is longer than the single method. The main reason is that the original data has to be decomposed into several components, and multiple predictors have to be built for each component. Therefore, the running time of hybrid-based CEEMDAN methods is generally longer.
4. Compared with CEEMDAN-LSTM, our proposed method performed faster and achieved higher prediction accuracy. Therefore, our proposed method is a suitable tool for predicting building energy consumption. Besides, the environment used in this study was CPU-based. With the development of a GPU-computing environment, the running time of deep learning methods such as LSTM could be enhanced up to 45 times faster than a single-threaded CPU implementation [72]. Thus, with the use of GPU, which can accelerate the computation, the running time of the proposed method can be further reduced.

6. Conclusion

Building energy consumption prediction is important for improving decision-making to achieve greater energy efficiency in the building. In this study, a hybrid Random Forest (RF) – Long Short-Term Memory (LSTM) based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is introduced to predict the hourly energy consumption in the building. In the first stage, the original energy consumption data were decomposed using CEEMDAN into several components. Then, the highest frequency component, which is the first component, was modeled using RF. The other components were predicted using LSTM. In the final stage, the prediction results of each component were integrated to obtain the final prediction result. By taking into account the characteristics of each component and utilizing different prediction methods to predict different components, the corresponding prediction result could be further enhanced. Hourly energy consumption data from five different buildings were used to evaluate the effectiveness of the proposed method. All experimental results indicated that the proposed method produced better results compared with other benchmarking methods. In this study, the prediction of energy consumption data was predicted based on its previous energy values (univariate prediction method). For future work, we will consider some exogenous variables in the prediction, such as weather conditions, building operational schedule, and time index (such as day of the week and hour of the day). We will also investigate the impact of different data set sizes with

different levels of granularity (such as daily and minutely data) on prediction performance.

7. Code availability

The source codes are made publicly available on Zenodo [73].

Declaration of Competing Interest

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

Acknowledgement

This study was financially supported by the Taiwan Building Technology Center from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

References

[1] M. Santamouris, Energy Consumption and Environmental Quality of the Building Sector, in: Minimizing Energy Consumption, Energy Poverty and Global and Local Climate Change in the Built Environment: Innovating to Zero, 2019. https://doi.org/10.1016/b978-0-12-811417-9.00002-7.
[2] REN21 Secretariat, 2020 Global Status Report for Buildings and Construction Towards a zero-emissions, efficient and resilient buildings owards a zero-emissions, efficient and resilient buildings and construction sector, 2020.
[3] IEA, Perspectives for a Clean Energy Transition. The Critical Role of Buildings., Energy Transition Progress and Outlook to 2020. (2019).
[4] A. Kalua, Urban residential building energy consumption by end-use in Malawi, Buildings 10 (2) (2020) 31, https://doi.org/10.3390/buildings10020031.
[5] Y. Ye, W. Zuo, G. Wang, A comprehensive review of energy-related data for U.S. commercial buildings, Energy and Buildings 186 (2019) 126–137, https://doi.org/10.1016/j.enbuild.2019.01.020.
[6] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205, https://doi.org/10.1016/j.rser.2017.04.095.
[7] BP, BP Energy Outlook 2019 edition, 2019.
[8] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy and Buildings 40 (3) (2008) 394–398, https://doi.org/10.1016/j.enbuild.2007.03.007.
[9] F. Qian, W. Gao, Y. Yang, D. Yu, Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption, Energy 193 (2020) 116724, https://doi.org/10.1016/j.energy.2019.116724.
[10] I. Antonopoulos, V. Robu, B. Couraud, D. Kirli, S. Norbu, A. Kiprakis, D. Flynn, S. Elizondo-Gonzalez, S. Wattam, Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review, Renewable and Sustainable Energy Reviews 130 (2020) 109899, https://doi.org/10.1016/j.rser.2020.109899.
[11] M. Ilbeigi, M. Ghomeishi, A. Dehghanbanadaki, Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm, Sustainable Cities and Society 61 (2020) 102325, https://doi.org/10.1016/j.scs.2020.102325.
[12] F. Pallonetto, C. Jin, E. Mangina, Forecast electricity demand in commercial building with machine learning models to enable demand response programs, Energy and AI 7 (2022) 100121, https://doi.org/10.1016/j.egyai.2021.100121.
[13] C. Kuster, Y. Rezgui, M. Mourshed, Electrical load forecasting models: A critical systematic review, Sustainable Cities and Society 35 (2017) 257–270, https://doi.org/10.1016/j.scs.2017.08.009.
[14] C. Liu, B. Sun, C. Zhang, F. Li, A hybrid prediction model for residential electricity consumption using holt-winters and extreme learning machine, Applied Energy 275 (2020) 115383, https://doi.org/10.1016/j.apenergy.2020.115383.
[15] P.-H. Kuo, C.-J. Huang, A high precision artificial neural networks model for short-Term energy load forecasting, Energies 11 (1) (2018) 213, https://doi.org/10.3390/en11010213.
[16] H. Dagdougui, F. Bagheri, H. Le, L. Dessaint, Neural network model for short-term and very-short-term load forecasting in district buildings, Energy and Buildings 203 (2019) 109408, https://doi.org/10.1016/j.enbuild.2019.109408.
[17] T. Catalina, V. Iordache, B. Caracaleanu, Multiple regression model for fast prediction of the heating energy demand, Energy and Buildings 57 (2013) 302–312, https://doi.org/10.1016/j.enbuild.2012.11.010.
[18] P. Ji, D. Xiong, P. Wang, J. Chen, A study on exponential smoothing model for load forecasting, in: Asia-Pacific Power and Energy Engineering Conference, APPEEC, 2012, https://doi.org/10.1109/APPEEC.2012.6307555.
[19] M.T. Hagan, S.M. Behr, The Time Series Approach to Short Term Load Forecasting, IEEE Transactions on Power Systems 2 (3) (1987) 785–791, https://doi.org/10.1109/TPWRS.1987.4335210.
[20] Z. Wang, Y. Wang, R. Zeng, R.S. Srinivasan, S. Ahrentzen, Random Forest based hourly building energy prediction, Energy and Buildings 171 (2018) 11–25, https://doi.org/10.1016/j.enbuild.2018.04.008.
[21] A. Lahouar, J. Ben Hadj Slama, Day-ahead load forecast using random forest and expert input selection, Energy Conversion and Management 103 (2015) 1040–1051, https://doi.org/10.1016/j.enconman.2015.07.041.
[22] M. Yalcintas, S. Akkurt, Artificial neural networks applications in building energy predictions and a case study for tropical climates, International Journal of Energy Research 29 (10) (2005) 891–901, https://doi.org/10.1002/er.1105.
[23] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545–553, https://doi.org/10.1016/j.enbuild.2004.09.009.
[24] B.-J. Chen, M.-W. Chang, C.-J. Lin, Load forecasting using support vector machines: A study on EUNITE Competition 2001, IEEE Transactions on Power Systems 19 (4) (2004) 1821–1830, https://doi.org/10.1109/TPWRS.2004.835679.
[25] J. Massana, C. Pous, L. Burgas, J. Melendez, J. Colomer, Short-term load forecasting in a non-residential building contrasting models and attributes, Energy and Buildings 92 (2015) 322–330, https://doi.org/10.1016/j.enbuild.2015.02.007.
[26] H. Choi, S. Ryu, H. Kim, Short-Term Load Forecasting based on ResNet and LSTM, in: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2018, 2018. https://doi.org/10.1109/SmartGridComm.2018.8587554.
[27] T.-Y. Kim, S.-B. Cho, Predicting residential energy consumption using CNN-LSTM neural networks, Energy 182 (2019) 72–81, https://doi.org/10.1016/j.energy.2019.05.230.
[28] W. Kong, Z.Y. Dong, Y. Jia, D.J. Hill, Y. Xu, Y. Zhang, Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network, IEEE Transactions on Smart Grid 10 (1) (2019) 841–851, https://doi.org/10.1109/TSG.2017.2753802.
[29] D. Liu, Q. Yang, F. Yang, Predicting Building Energy Consumption by Time Series Model Based on Machine Learning and Empirical Mode Decomposition, in: 2020 5th IEEE International Conference on Big Data Analytics, ICBDA 2020, 2020, https://doi.org/10.1109/ICBDA49040.2020.9101335.
[30] H. Zheng, J. Yuan, L. Chen, Short-Term Load Forecasting Using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation, Energies 10 (8) (2017) 1168, https://doi.org/10.3390/en10081168.
[31] N. An, W. Zhao, J. Wang, D. Shang, E. Zhao, Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting, Energy 49 (2013) 279–288, https://doi.org/10.1016/j.energy.2012.10.035.
[32] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, N. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proceedings of the Royal Society A. 454 (1996).
[33] D. Zhang, C. Cai, S. Chen, L. Ling, An improved genetic algorithm for optimizing ensemble empirical mode decomposition method, Systems Science and Control Engineering 7 (2) (2019) 53–63, https://doi.org/10.1080/21642583.2019.1627598.
[34] Z. Wu, N.E. Huang, Ensemble empirical mode decomposition: A noise-assisted data analysis method, Advances in Adaptive Data Analysis 01 (01) (2009) 1–41, https://doi.org/10.1142/S1793536909000047.
[35] J. Zheng, J. Cheng, Y. Yang, Partly ensemble empirical mode decomposition: An improved noise-assisted method for eliminating mode mixing, Signal Processing 96 (2014) 362–374, https://doi.org/10.1016/j.sigpro.2013.09.013.
[36] X. Zhang, Y. Yang, Suspended sediment concentration forecast based on CEEMDAN-GRU model, Water Science and Technology: Water Supply. 20 (2020). https://doi.org/10.2166/ws.2020.087.
[37] M.E. Torres, M.A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2011. https://doi.org/10.1109/ICASSP.2011.5947265.
[38] H. Lin, Q. Sun, S.-Q. Chen, Reducing exchange rate risks in international trade: A hybrid forecasting approach of CEEMDAN and multilayer LSTM, Sustainability (Switzerland) 12 (6) (2020) 2451, https://doi.org/10.3390/su12062451.
[39] L. Breiman, Random forests. Mach Learn, Random Forests. (2001).
[40] L. Lin, F. Wang, X. Xie, S. Zhong, Random forests-based extreme learning machine ensemble for multi-regime time series prediction, Expert Systems with Applications 83 (2017) 164–176, https://doi.org/10.1016/j.eswa.2017.04.013.
[41] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (2002).
[42] A.J. Sage, Random Forest Robustness, Variable Importance, and Tree Aggregation, ProQuest Dissertations and Theses (2018).
[43] Z.H. Zhou, J. Feng, Deep forest: Towards an alternative to deep neural networks, in: IJCAI International Joint Conference on Artificial Intelligence, 2017. https://doi.org/10.24963/ijcai.2017/497.
[44] Y. Guo, S. Liu, Z. Li, X. Shang, BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data, BMC Bioinformatics 19 (S5) (2018), https://doi.org/10.1186/s12859-018-2095-4.

[45] G. Zhu, Q. Hu, R. Gu, C. Yuan, Y. Huang, ForestLayer: Efficient training of deep forests on distributed task-parallel platforms, Journal of Parallel and Distributed Computing 132 (2019) 113–126, https://doi.org/10.1016/j.jpdc.2019.05.001.
[46] R. Srivastava, A.N. Tiwari, V.K. Giri, Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India, Heliyon 5 (10) (2019) e02692, https://doi.org/10.1016/j.heliyon.2019.e02692.
[47] C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy consumption, Renewable and Sustainable Energy Reviews 74 (2017) 902–924, https://doi.org/10.1016/j.rser.2017.02.085.
[48] M. Elsaraiti, A. Merabet, Application of long-short-term-memory recurrent neural networks to forecast wind speed, Applied Sciences (Switzerland) 11 (5) (2021) 2387, https://doi.org/10.3390/app11052387.
[49] F. Cheng, J. Zhao, A novel process monitoring approach based on Feature Points Distance Dynamic Autoencoder, Computer Aided Chemical Engineering (2019), https://doi.org/10.1016/B978-0-12-818634-3.50127-2.
[50] S. Hochreiter, J. Schmidhuber, Long short term memory. Neural computation, Neural Computation 9 (8) (1997) 1735–1780.
[51] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural Computation 31 (7) (2019) 1235–1270, https://doi.org/10.1162/neco_a_01199.
[52] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, M. Weyrich, A survey on long short-term memory networks for time series prediction, Procedia CIRP 99 (2021) 650–655, https://doi.org/10.1016/j.procir.2021.03.088.
[53] H. Yan, Y. Qin, S. Xiang, Y. Wang, H. Chen, Long-term gear life prediction based on ordered neurons LSTM neural networks, Measurement: Journal of the International Measurement Confederation 165 (2020) 108205, https://doi.org/10.1016/j.measurement.2020.108205.
[54] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP 2014–2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, https://doi.org/10.3115/v1/d14-1179.
[55] X. Wang, J. Xu, W. Shi, J. Liu, OGRU: An Optimized Gated Recurrent Unit Neural Network, Journal of Physics: Conference Series 1325 (1) (2019) 012089, https://doi.org/10.1088/1742-6596/1325/1/012089.
[56] M. Zekić-Sušac, A. Has, M. Knežević, Predicting energy cost of public buildings by artificial neural networks, CART, and random forest, Neurocomputing 439 (2021) 223–233, https://doi.org/10.1016/j.neucom.2020.01.124.
[57] N. Zhang, Z. Li, X. Zou, S.M. Quiring, Comparison of three short-term load forecast models in Southern California, Energy 189 (2019) 116358, https://doi.org/10.1016/j.energy.2019.116358.
[58] J. Xu, K. Xu, Z. Li, F. Meng, T. Tu, L. Xu, Q. Liu, Forecast of dengue cases in 20 chinese cities based on the deep learning method, International Journal of Environmental Research and Public Health 17 (2) (2020) 453, https://doi.org/10.3390/ijerph17020453.
[59] S.-L. Lin, H.-W. Huang, Improving Deep Learning for Forecasting Accuracy in Financial Data, Discrete Dynamics in Nature and Society 2020 (2020) 1–12, https://doi.org/10.1155/2020/5803407.
[60] N. Halpern-Wight, M. Konstantinou, A.G. Charalambides, A. Reinders, Training and testing of a single-layer LSTM network for near-future solar forecasting, Applied Sciences (Switzerland) 10 (17) (2020) 5873, https://doi.org/10.3390/app10175873.
[61] C. Miller, F. Meggers, The Building Data Genome Project: An open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444, https://doi.org/10.1016/j.egypro.2017.07.400.
[62] D. Laszuk, PyEMD Documentation, (2020).
[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011).
[64] F. Chollet, Keras Documentation, Keras.Io. (2015).
[65] A. Lahouar, J. Ben Hadj Slama, Hour-ahead wind power forecast based on random forests, Renewable Energy 109 (2017) 529–541, https://doi.org/10.1016/j.renene.2017.03.064.
[66] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.
[67] A. Kumar Dubey, A. Kumar, V. García-Díaz, A. Kumar Sharma, K. Kanhaiya, Study and analysis of SARIMA and LSTM in forecasting time series data, Sustainable Energy Technologies and Assessments 47 (2021) 101474, https://doi.org/10.1016/j.seta.2021.101474.
[68] A. Kara, Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm, Expert Systems with Applications 180 (2021) 115153, https://doi.org/10.1016/j.eswa.2021.115153.
[69] C. Jörges, C. Berkenbrink, B. Stumpe, Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks, Ocean Engineering 232 (2021) 109046, https://doi.org/10.1016/j.oceaneng.2021.109046.
[70] I. Kandel, M. Castelli, The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset, ICT Express 6 (4) (2020) 312–315, https://doi.org/10.1016/j.icte.2020.04.010.
[71] Y. Wang, J. Wang, Z. Li, A novel hybrid air quality early-warning system based on phase-space reconstruction and multi-objective optimization: A case study in China, Journal of Cleaner Production 260 (2020) 121027, https://doi.org/10.1016/j.jclepro.2020.121027.
[72] I.M. Coelho, V.N. Coelho, E.J.da.S. Luz, L.S. Ochi, F.G. Guimarães, E. Rios, A GPU deep learning metaheuristic based model for time series forecasting, Applied Energy 201 (2017) 412–418, https://doi.org/10.1016/j.apenergy.2017.01.003.
[74] irenekarijadi. (2021). irenekarijadi/RF-LSTM-CEEMDAN: (v2.0). Zenodo. doi: 10.5281/zenodo.5930048..