

Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion

Jinsong Wang, Xuhui Chen, Fan Zhang, Fangxi Chen, and Yi Xin

Abstract—The energy consumption of buildings has risen steadily in recent years. It is vital for the managers and owners of the building to manage the electric energy demand of the buildings. Forecasting electric energy consumption of the buildings will bring great profits, which is influenced by many factors that make it very difficult to provide an advanced forecasting. Recently, deep learning techniques are widely adopted to solve this problem. Deep neural network offers an excellent capability in handling complex non-linear relationships and competence in exploring regular patterns and uncertainties of consumption behaviors at the building level. In this paper, we propose a deep convolutional neural network based on ResNet for hour-ahead building load forecasting. In addition, we design a branch that integrates the temperature per hour into the forecasting branch. To enhance the learning capability of the model, an innovative feature fusion is presented. At last, sufficient ablation studies are conducted on the point forecasting, probabilistic forecasting, fusion method, and computation efficiency. The results show that the proposed model has the state-of-the-art performance, which reflects a promising prospect in application of the electricity market.

Index Terms—Load forecasting, deep learning, convolutional neural network, feature fusion, ResNet.

I. INTRODUCTION

INCREASING energy demands have attracted more attention with the growth of the world population and economic development. At the same time, in order to solve problems of the pollution, carbon emissions and greenhouse, reducing energy consumption must be taken into account. According to the U.S. Energy Information Administration Monthly Energy Review, 40% of the energy consumption comes from buildings [1]. In addition, with the aim of achieving the building energy conservation, some policies and regulations have been promulgated for the effective design of new buildings in many countries. The growing energy demand of the buildings requires reliable load forecasting, which will

promote the effective planning, long-term strategies, effective plans to reduce carbon emissions, and control energy usage in the construction sector [2]. A great number of innovative techniques have been introduced for smart grids to improve the power system reliability and building energy efficiency, including demand response (DR) [3] and demand-side management [4], [5]. From the perspective of electric operators, accurate building load forecasting ensures the effectiveness of both pre-DR resource assignment and post-DR performance assessment [6]. Modeling and forecasting the energy consumption of buildings are essential for urban areas to reduce their overall energy consumption [7]. Reasonable consumption forecasting becomes significant in that it could save 10% to 30% of the building energy consumption [8]. Energy consumption forecasting is an important part of the energy management system. It aims to provide the key information for daily management and grid planning of electric utility, which make optimal decisions in grid energy management to ensure safe and reliable operation of power system. It is proved that improving the energy efficiency of buildings by designing accurate and powerful load forecasting models is an effective solution for energy management, DR procedures, fault detection, and energy benchmarking [9].

Since the energy and building types are various, the energy system in buildings is quite complicated [10]. Most energy is consumed by heating, ventilating, and air-conditioning (HVAC) system, water heater, and other electric appliances. Common building types consist of residential, office, and engineering buildings. The energy consumption of buildings is influenced by internal and external factors. Internal factors include sub-level components such as lighting, HVAC systems, and the occupancy behavior. External ones such as weather conditions, and thermal property of the used physical materials also affect electric demands. Due to the regular pattern and uncertainty in building load profiles, caused by internal and external factors, it is difficult to make a precise short-term load forecasting.

In recent years, a large number of forecasting models have been proposed and applied to solve practical problems. The load forecasting category can be simply summarized as follows: very short-term load forecasting (VSTLF), short-term load forecasting (STLF), medium-term load forecasting (MTLF), and long-term load forecasting (LTLF). The main forecasting models are divided into two categories: traditional and artificial intelligence methods. The advantages of tra-

Manuscript received: May 21, 2020; accepted: September 3, 2020. Date of CrossCheck: September 3, 2020. Date of online publication: January 11, 2021.

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

J. Wang (corresponding author) and F. Zhang are with the Department of Electrical, Computer, and System Engineering at Case Western Reserve University, Cleveland, USA (e-mail: jxw741@case.edu; fan.zhang@case.edu).

X. Chen is with the College of Aeronautics and Engineering, Kent State University, Kent, USA (e-mail: xchen58@kent.edu).

F. Chen and Y. Xin are with Software College, Northeastern University, Shenyang, China (e-mail: 924290841@qq.com; xiny0919@163.com).

DOI: 10.35833/MPCE.2020.000321



ditional methods depend on their clipping computing speeds and robustness [11], which include linear regression (LR), multiple linear regression (MLR), and auto-regressive integrated moving average (ARIMA). Reference [12] studies the sensitivity of Iran electrical load to the temperature based on the LR method. The analysis can be extended to all other environmental factors such as humidity, wind speed, and weather coverage and can also be carried out on the power grids of other countries. Reference [13] proposes a modern processing method of classical technology, MLR, to model hourly demand and study the causality of power consumption. The proposed model has been used to generate a 3-year hourly energy demand forecasting for the U.S. companies. Reference [14] combines several methods of ARIMA based on the idea of time series to avoid the shortcomings in various aspects, helping ARIMA model better forecast the short-term load. Reference [15] depends on the lifting scheme and ARIMA model. The lifting scheme is a general and flexible method for constructing biorthogonal wavelets, which are embedded in the ARIMA model to improve the forecasting accuracy. Based on the results of wavelet multiple resolution analysis (MRA), the lifting scheme decomposes the original load sequence into subsequences with different resolutions. Then, the inverse lifting is used to reconstruct the forecasting results of different levels to generate the original load forecasting.

Support vector machine (SVM) is a kernel-based machine learning algorithm, which can be used for both regression and classification. The algorithm reveals great competence in non-linear analysis. Support vector regression (SVR) [16]-[19] is an important application branch of SVM. With the kernel function, training data are potentially transformed into higher dimensional spaces, which benefits extracting discriminative features and strengthens the capability of model. It has been successfully used in solving nonlinear regression and time series problems in the aspects of building load forecasting. Since the performance of SVR heavily relies on the selection of its parameters, [20] uses the differential evolution (DE) algorithm to solve this problem. The forecasting model is developed using a weighted SVR model with nu-SVR and epsilon-SVR, and DE determines the weight corresponding to each model. The proposed model can be used to forecast half-hourly and daily electricity time series data of the same building. However, SVR is not stable enough for outliers, and the setting of training parameters involves many techniques and difficulties, which leads to a poor training process.

Strategies based on machine learning offer a restricted capability of modeling in time series analysis. In recent years, neural networks have appeared as a powerful tool in the areas such as image processing and data analysis. Deep learning skills have been applied in load forecasting, and most of them rely on recurrent neural network (RNN) or long short-term memory (LSTM) [21]-[24]. LSTM is derived from RNN, and both of them are successful in the target of sequence-to-sequence learning such as speech recognition and natural language processing in time-series analysis. Gated re-

current neural (GRU) network is an effective variant of LSTM and has a simpler structure. It can also solve the problem of long dependency in RNN network. Therefore, it is also a very manifold network at present. However, when managing long-term sequence, RNNs suffer from the problem of gradient disappearance severely, even though LSTM alleviates this case partly. Moreover, the RNN-based models require serial calculation, so the calculation efficiency is unsatisfactory.

Convolutional neural networks (CNNs) have achieved brilliant performances in the field of computer vision with well-known models such as AlexNet, ResNet, and DenseNet. ResNet was proposed in 2015, and it is the most widely popular CNN framework for feature extraction at present. Recently, ResNet models also demonstrate better performances in sequence processing, not only in speech synthesis, language modeling and machine translation, but also in electricity load forecasting. ResNet model contains a number of residual connections between different level blocks and delivers the errors to previous layers when the network is being trained with back propagation. This mechanism is able to increase the depth of the network and strengthen the ability to learn discriminative features. In other words, ResNet can effectively deal with the problem of gradient disappearance. An example of ResNet structure is illustrated in Fig. 1, where the input x is the sequence of historical data; W_1 and W_2 are the weights; and $F(x)$ is the output of two convolutional layers.

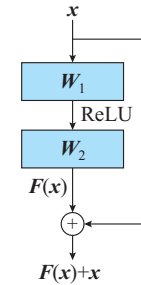


Fig. 1. Structure of ResNet.

With a deeper architecture and sophisticated operation, neural networks, especially deep CNNs (DCNNs), provide superior abilities of learning discriminative features and non-linear relationships, which benefits extracting uncertainty factors of building load forecasting [25]-[27]. Specifically, the latest researches [26], [27] reveal that CNN performs a more advanced accuracy as a result of the powerful capability of the discriminative feature extraction. Gated CNN (GCNN) is a deep learning model. GCNN introduces the gated mechanism of LSTM into the CNN and uses the gated mechanism to identify and judge the information, which has achieved superior results. In addition, some mechanisms like residual connection cannot cause dramatical gradient disappearance even in deeper network. Consequently, related skills could be optimized to identify and learn both the regular pattern and uncertainty in load profiles. DCNN performs superiorly in estimating heating/cooling loads, total electricity consump-

tion, and operation and optimization of sub-level components.

Although the range of forecasting time may vary from a few minutes to several years, especially shorter than a day, it is critical for buildings because utility prices may vary with the season and time. A more effective and efficient estimation of the peak electricity load in a day and the load shape has more possibilities to control the utility costs for providing more intelligence in smart buildings. In this paper, hourly electricity consumption data and temperature data are adopted to forecast hourly consumption one hour ahead (single-step forecasting) and 24 hours ahead (24-step forecasting).

Based on the literature outlined above, this paper proposes a novel CNN model for sub-hourly building load forecasting. The key contributions are as follows.

1) We propose a novel DCNN for building load forecasting. The baseline of the network is built on ResNet containing a number of residual connections in order to increase the depth of the model and enhance the ability of learning non-linear relationships in time-series analysis.

2) For integrating the information of external factors into the forecasting model, we design another branch responsible for extracting the pattern of external factors per hour with fully connected layers.

3) A novel mechanism of feature fusion between two branches is proposed, which is also interpreted as a superior feature selection process and leads to a remarkable improvement in accuracy.

4) The proposed network serves in an end-to-end manner during the training and inference process. Sufficient ablation studies are conducted to demonstrate the effectiveness of innovations and great generalization in building load forecasting.

The rest of this paper will be organized as follows. Section II describes the details of the proposed model. Section III introduces the elements and details of the experiment process. Section IV reports the experimental results and discusses some doubts of the proposed model. Finally, the conclusions are summarized in Section V.

II. METHODOLOGY

A. Problem Formulation

This paper proposes a CNN model with two branches including the forecasting branch and the external factor integration branch. There is an input \mathbf{x}^L of main network representing the sequence of historical load, and another input \mathbf{x}^W of branch represents outdoor weather variables. The vector \mathbf{Y} is the output forecasting. The main purpose is to build a corresponding mapping relationship between inputs and future load sequences as follows.

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{e} \quad (1)$$

where $\mathbf{X} = [x_1, x_2, \dots, x_{M_a}]$, $\mathbf{Y} = [y_{M_a+1}, y_{M_a+2}, \dots, y_{M_a+M_b}]$, x_1, x_2, \dots, x_{M_a} are the historical loads, $y_{M_a+1}, y_{M_a+2}, \dots, y_{M_a+M_b}$ are the output forecasting, and M_a and M_b are the lengths of the input

and output sequences, respectively; and \mathbf{e} is the error. When M_b is 1, it is a single-step forecasting problem, and if M_b is over 1, it is a multi-step forecasting problem.

B. Network Architecture

In this paper, building load forecasting is solved as a sequence issue. Deep networks based on DCNN automatically extract the key change features layer by layer in the historical load sequence, and generate forecasting at the end of the model. External factors such as temperature are also the key features in building load forecasting and should be considered in the model. Changes of external factors will affect the pattern of load sequences. This effect is broad and comprehensive. For example, in hot weather, due to the use of air conditioning, the load changes more drastically during the day; on holidays and working days, the load curve changes quite differently. The external factors used in this paper refer specifically to historical data, since the focus here is on the complex relationship between historical external variables and load data.

Therefore, we propose a novel deep network structure to fuse sequence change features and external factor features. In this network, the external factor features are explicitly modeled as the constraints of the historical load, constraining the original input and changing characteristics of the historical load. Specifically, the learning to external variables generates indicator vectors, where each scalar is close to 0 or 1. Element-wise multiplication is done based on indicators with different layers in DCNN to control the expression of the key features. This is very different from the traditional feature fusion (TFF) ways, which generally concatenate sequence change features and external factor features, and generate forecasting through fully connected layers. Our model explicitly models the relationship between sequence change features and external factor features, introducing strong priors to the model. Therefore, this network could converge to a more optimal local minimum, which improves the forecasting accuracy.

Both of the forecasting branch and external factor integration branch run in a parallel way and constitute an end-to-end manner of training and inference. Forecasting branch serves for building load forecasting and external factor integration branch takes responsibilities of increasing additional significant features for efficient and effective forecasting exploring potential nonlinear relationships. Details of our proposed network are illustrated in Fig. 2.

C. Forecasting Branch

The forecasting branch is a CNN that aims to extract the features of electric load sequences. It consists of a series of stacked residual blocks. In this study, we employ 4 blocks, each of which is composed of four layers with different functions: dilated convolution, Relu activation, normalization, and regular 1-dimension (1D) convolution, respectively, as shown in Fig. 3. The forecasting branch receives a 24-dimension load vector into the network. In addition, five residual blocks join in the structure of this branch.

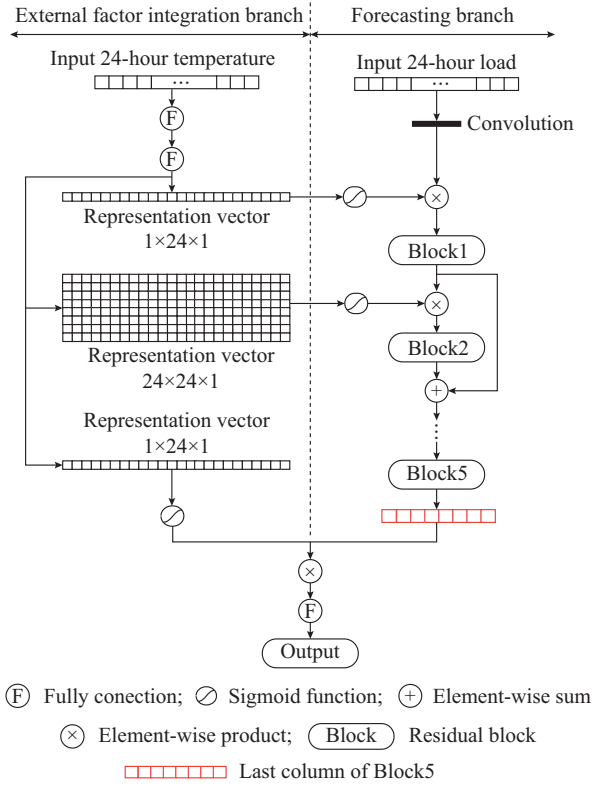


Fig. 2. Architecture of our proposed model.

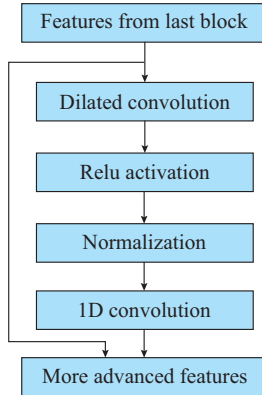


Fig. 3. Illustration of one residual block as baseline of proposed forecasting branch.

Dilated convolution focuses on extracting a larger-scale of features in local receptive field and results in fusion of different dilated ratios, which means that comprehensive feature abstracting leads to an advanced performance in deep learning. Dilated convolution is widely used in image segmentation which can greatly increase the receptive field without increasing the computation complexity and the number of parameters. Therefore, we introduce the dilated convolution to extend receptive fields of a neuron by embedding zero-value holes at various scales. Dilated convolution does not mean that the blank elements are padded between the elements of the feature map when convoluted, but some elements are skipped over the existing elements, or the input is unchanged in this way, and some zero weights are inserted into kernel convolution parameters. In addition, dilated con-

volution introduces a parameter called dilation rate to indicate the expansion rate. In this study, we adopt the dilated convolution with rates 1, 2, 4, and 8, respectively. We also introduce residual connections to ensure that the gradient of our model will not disappear or explode due to the depth. Leaky Relu activation is responsible to filter salient features for data analysis [28]. Normalization strategy introduces batch normalization (BN) [29] that preserves identical distribution in our dataset and avoids from gradient explosion in the training process. The last component, regular 1D convolution, pays more attention to extracting patterns in neighborhoods. Every block is adopted for collecting more advanced features from the outputs of the previous block.

D. External Factor Integration Branch

Building load forecasting cannot live without external factors, especially weather conditions that have been proven a strong correlation in this field. External factor integration branch takes responsibilities of increasing additional significant features to improve the accuracy of building load forecasting. This branch contains multiple fully-connected layers with different numbers of activation functions in order to produce feature vectors with proper dimensions.

Specifically, the external factor integration branch takes a 24-dimension vector as the input, where each element represents the historical external factors per hour, totally 24 hours. Then, two fixed sizes of fully connected layers filter the input to extract coarse nonlinear features. Moreover, there are three hidden layers derived as outputs that are fused to forecasting branch, $1 \times 24 \times 1$, $24 \times 24 \times 1$, and $1 \times 24 \times 1$, respectively, as illustrated in Fig. 2. All activation functions adopt Sigmoid function, which ensures the values of trainable parameters fall within (0, 1). The reason for selecting these output dimensions of the three hidden layers is to keep consistent with the output dimensions of the corresponding layer in the forecasting branch, so as to facilitate element-wise multiplication (i.e., feature fusion). The external factor integration branch realizes the extracting of external factors in the forecasting branch according to the external environment.

E. Feature Fusion

CNN extracts the changing features in the load sequence and generates feature maps. Although deeper convolution layers can simultaneously extract and select features in feature maps, this process does not consider external factors.

In this paper, we propose a novel feature fusion or feature selection process shown in Fig. 2, where outputs of the external factor integration branch with learnable weights are fused into the baseline of forecasting branch by multiplication operation. Three outputs of household profile branch are set to be $1 \times 24 \times 1$, $24 \times 24 \times 1$, and $1 \times 24 \times 1$, respectively. In the forecasting branch, we choose the input layer, the first convolution block, and the last convolution block for fusion, which represent the input, low-level features, and high-level features, respectively. When the input layer is constrained, this model is similar to the original models for feature selection of regression except that it changes along with external factors. When the convolution layer is constrained, this mod-

el is similar to a gating model except that it uses external variables. In general, element-wise multiplication fusion ensures a feature selection process when the vector from building profile branch is filtered by Sigmoid activation and elements are fallen within (0, 1). As a result of the superior mechanism in weight learning, the external factor integration branch provides an excellent encoding of external factors, which selects salient features in different levels of forecasting branch. Consequently, most significant features are delivered to the next block, giving contributions to final fully-connected layers of the entire network. The two branches form an end-to-end manner for training and inference with a parallel approach. Therefore, the entire model is able to explore more non-linear relationships among consumption behaviors of buildings, achieving more competent performance with great generalization for load forecasting.

III. EXPERIMENT SETUP

A. Data Description

We adopt the dataset from the genome project building data [30], which includes 507 public datasets from electrical meters of non-residential building. Each dataset includes load and the corresponding weather conditions. We choose two laboratory buildings and an office building as the research objects, denoted as buildings *A*, *B*, and *C*, in Switzerland, respectively. The information of the three buildings is shown in the Table I. The area of building *A* is larger than others. The time starts from January 1, 2013 to December 31, 2013. Peak electricity demands of buildings *A*, *B*, and *C* are approximately 90, 85, and 25 kW/h in summer, which are higher than those in winter. The national average annual temperature in Switzerland is 8.6 °C. In summer, the average temperature is 18 °C to 27 °C (rarely above 30 °C), and the temperatures of the day and night are greatly different. Average temperature in winter is −1 °C to −5 °C, relatively not very cold.

TABLE I
DETAILS OF 3 BUILDINGS FOR EVALUATIONS IN ABLATION STUDY

Building	Type	Area (m ²)	Location
<i>A</i>	Laboratory	6875	Zurich
<i>B</i>	Laboratory	6039	Zurich
<i>C</i>	Office	186	Zurich

In addition, more buildings are added to our experiments to verify the forecasting ability of the proposed model. After eliminating the abnormal data, we adopt the data of a total of 300 buildings from January 1, 2010 to December 30, 2015, which are located in New York, Los Angeles, Chicago, Phoenix, London, and Switzerland, respectively. The buildings are all used for education, including offices, dormitories, laboratories, and classrooms. The floor areas range from 399 to 155679 m², and their loads range from 1 to 823 kW.

B. Weather-relevant Feature Selection

In order to select more appropriate weather variables as the input of external factor integration branch, it is vital to

select the most relevant features for building load forecasting before the model construction. There are some weather variables considered as candidates, which includes outdoor temperature (\mathbf{x}^T), humidity (\mathbf{x}^H), and wind speed (\mathbf{x}^S). Pearson correlation coefficient is also called simple correlation coefficient, which describes the closeness and correlation of the relationship between two variables. Its value stays between −1 and 1. The value of 1 indicates that the variable is completely positively correlated, 0 expresses irrelevant, and −1 means completely negatively correlated. Its calculation process can be summarized as:

$$P_{\mathbf{x}^W, \mathbf{x}^L} = \frac{\text{cov}(\mathbf{x}^W, \mathbf{x}^L)}{\omega_{\mathbf{x}^W} \omega_{\mathbf{x}^L}} \quad \mathbf{x}^W \in \{\mathbf{x}^T, \mathbf{x}^H, \mathbf{x}^S\} \quad (2)$$

where $\text{cov}(\cdot)$ represents the covariance; and $\omega_{\mathbf{x}^W}$ and $\omega_{\mathbf{x}^L}$ are the standard deviations of any weather variable \mathbf{x}^W and building load series \mathbf{x}^L , respectively.

Table II summarizes the Pearson correlation coefficients statistics between each weather variable and building load for buildings *A*, *B*, and *C*, respectively. As is shown in Table II, there is a stronger positive correlation between \mathbf{x}^T and \mathbf{x}^L for all three buildings in comparison with other weather variables and building load. Adding weakly correlated variables into our model will introduce unnecessary noises. Therefore, in this study, only \mathbf{x}^T is selected as the weather-relevant feature.

TABLE II
PEARSON CORRELATION COEFFICIENTS BETWEEN EACH WEATHER VARIABLE AND BUILDING LOAD

Building	$P_{\mathbf{x}^T, \mathbf{x}^L}$	$P_{\mathbf{x}^H, \mathbf{x}^L}$	$P_{\mathbf{x}^S, \mathbf{x}^L}$
<i>A</i>	0.75	−0.21	0.03
<i>B</i>	0.69	−0.09	0.05
<i>C</i>	0.43	−0.14	0.16

C. Data Preprocessing

The frequency of dataset acquisition is once an hour, and consequently there are 24 data points per day. The raw data may have noisy, missing or redundant variables. Therefore, raw data are preprocessed to ensure the dataset availability before experiments. The datasets are divided into training, validation, and testing dataset by 80%, 10%, and 10% in chronological order, respectively. Finally, in order to stabilize the learning process, input variables with corresponding validation and testing datasets are meticulously normalized. Normalization helps avoid dramatic changes on the gradient, which is beneficial to smoothen the convergence.

After all datasets are properly preprocessed, model parameters, i.e., weights and bias, and hyper-parameters, i.e., layers and number of neurons, are tuned using training and validation datasets. Once the optimal model parameters and hyper-parameters are obtained, the testing dataset of each building would be fed into the optimized models to evaluate performances.

D. Selection of Benchmark and Hyperparameter

In this study, our model is evaluated in comparison with GRU, LSTM, GCNN, and ResNet. Details of parameters are shown in Table III. Hyperparameters of LSTM keeps the

same with GRU. Hyperparameters of GCNN is set as those in [31].

TABLE III
PARAMETERS OF PROPOSED MODEL AND OTHER MODELS

Model	Depth	Kernel size	Kernel number	Batch size	Loss function	Optimizing model	Learning rate	Training stop
ResNet	34	8	24	128	Mean square error (MSE)	AMSGrad	0.002	Early stopping
GRU/LSTM	3	None	None	128	MSE	AMSGrad	0.100	Early stopping
Proposed model	8	8	24	128	MSE	AMSGrad	0.002	Early stopping

For fair comparison, some setup and hyperparameters of all models should be as consistent as possible, such as the batch size, optimization method, and number of trainable parameters. On this basis, we adopt the grid search method to determine other hyperparameters, including the learning rate, network depth, convolution kernel, and so forth.

To ensure the fairness, historical temperature data are also considered in four benchmark models. Unlike the proposed model, which handles the temperature and load sequences in two branches, historical temperature and load data in the benchmark models are integrated as two dimensions of the input sequence. The input vector form of the benchmark models is:

$$\mathbf{X} = \begin{bmatrix} x_1^L & x_2^L & \dots & x_{24}^L \\ x_1^T & x_2^T & \dots & x_{24}^T \end{bmatrix} \quad (3)$$

E. Software and Hardware Platform

All experiments are conducted on a cloud server with the CPU with 8 cores and 2 NVIDIA P4 computing cards. Neural network based models are realized by the Keras framework with Tensorflow backend.

IV. RESULT AND DISCUSSION

A. Evaluation Metric

We carry out point and probabilistic forecasting, separately, and evaluate the point forecasting results of the three buildings by the following three metrics, including mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE). Smaller values from these metrics mean that the model has lower errors and higher accuracies. The metrics are calculated based on (4)-(6).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (5)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{y}_i - y_i|}{y_i} \times 100 \quad (6)$$

where \hat{y}_i and y_i are the forecasting and true values, respectively; and N is the amount of data.

In order to estimate the quality of probabilistic forecasting, we adopt a comprehensive metric called pinball score (also called quantile score). It is widely recognized and can be interpreted as the accuracy of a quantile forecasting model. Pinball score for one quantile can be calculated by the following formula:

$$P = \begin{cases} (1-q)(\hat{Y}_t^q - Y_t) & \hat{Y}_t^q \geq Y_t \\ q(Y_t - \hat{Y}_t^q) & \hat{Y}_t^q < Y_t \end{cases} \quad (7)$$

where q is the targeted quantile; and \hat{Y}_t^q and Y_t are the forecasting and true values in the q^{th} quantile at time t , respectively.

B. Point Forecasting

The point forecasting accuracies of the three buildings are presented in Tables IV, V, and VI, respectively. Our proposed network performs significantly better than other models, and the LSTM performs the worst. As shown in Table IV, the MAPE of our proposed model obviously decreases compared with the other three models for smaller buildings, as MAPE remains around 2%. Moreover, the MAEs of the proposed model on the three buildings are 24.3%, 43.6%, and 56.3% lower than the optimal MAE of other models, respectively. This indicates a much better generalizability of our model compared with GRU, ResNet, LSTM, and GCNN. The comparison of the results is shown in Fig. 4. The points of the proposed model are closer to the diagonal line, indicating that the forecasting accuracy is higher than those of several other models, respectively. The forecasting values of the proposed model are most closely distributed near the ground truth, revealing its accuracy is better than the other four models.

TABLE IV
RESULTS OF MAPE AND STANDARD DEVIATION FOR SINGLE-STEP POINT FORECASTING WITH FIVE MODELS

Building	GRU		ResNet		LSTM		GCNN		Proposed model	
	MAPE (%)	Standard deviation (%)	MAPE (%)	Standard deviation (%)	MAPE (%)	Standard deviation (%)	MAPE (%)	Standard deviation (%)	MAPE (%)	Standard deviation (%)
A	2.93	2.02	2.63	1.68	3.60	2.45	2.72	2.36	2.22	1.53
B	4.56	2.83	8.06	5.56	4.90	2.89	6.01	4.27	2.80	1.96
C	8.34	5.59	6.87	4.67	9.04	6.33	6.74	4.58	4.26	2.81

TABLE V
RESULTS OF MAE AND STANDARD DEVIATION FOR SINGLE-STEP POINT FORECASTING WITH FIVE MODELS

Building	GRU		ResNet		LSTM		GCNN		Proposed model	
	MAE (kW)	Standard deviation (kW)	MAE (kW)	Standard deviation (kW)	MAE (kW)	Standard deviation (kW)	MAE (kW)	Standard deviation (kW)	MAE (kW)	Standard deviation (kW)
<i>A</i>	2.57	1.69	1.95	1.20	2.60	1.74	1.89	1.19	1.43	0.96
<i>B</i>	2.89	1.64	4.00	2.73	3.25	1.88	3.13	2.13	1.63	1.12
<i>C</i>	4.08	2.60	3.00	1.87	4.40	2.83	2.86	1.78	1.25	0.78

TABLE VI
COMPARISON OF RMSE FOR SINGLE-STEP POINT FORECASTING WITH FIVE MODELS

Building	RMSE (kW)				
	GRU	ResNet	LSTM	GCNN	Proposed model
<i>A</i>	2.84	2.57	3.62	2.57	2.09
<i>B</i>	3.85	4.94	4.35	4.05	2.27
<i>C</i>	5.04	4.46	6.40	4.08	1.67

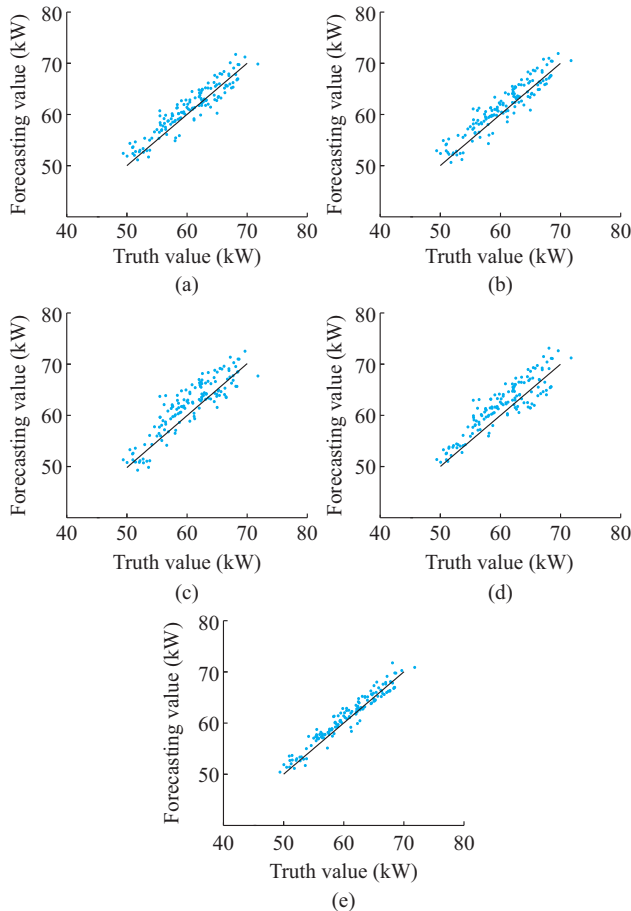


Fig. 4. Scatter plot of five models for building *A*. (a) ResNet. (b) GCNN. (c) LSTM. (d) GRU. (e) Proposed model.

Building *A* has lower MAPE values compared with the other two buildings. Along with the load decrease of different buildings, the accuracies of all models also go down (as shown from the results for buildings *B* and *C*). One of the

reasons is larger buildings usually accommodate more occupiers. When there are more occupiers in a building, the uncertainty of their overall behaviors tends to be smaller, leading to more regular and predictable building load patterns. Another reason is the Pearson correlation coefficient of building *A* is the highest, which makes its forecasting benefits more from the temperature.

The MAPE of a more extensive experiment on 300 buildings is shown in Fig. 5, where warmer color indicates lower forecasting errors. The line corresponding to the proposed model is significantly warmer than the other lines, illustrating a generally lower forecasting error. Compared with the other four models, the proposed model shows general improvements. The proposed model has an average MAPE reduction of 29.7%, 32.8%, 35.9%, and 25.3% relative to GRU, ResNet, LSTM, and GCNN, respectively.

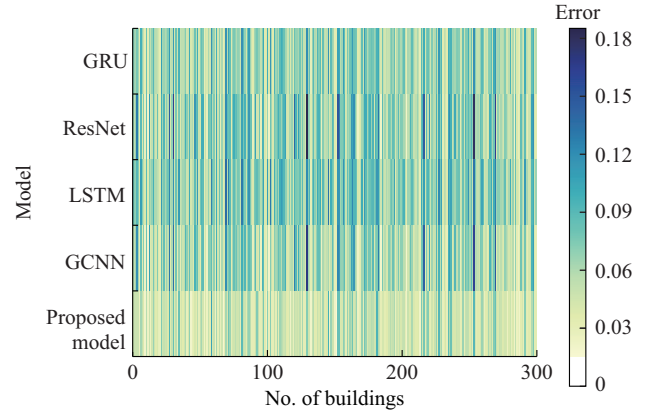


Fig. 5. MAPE heat map of single-step forecasting on 300 buildings.

In the practice of DR, 24-step forecasting provides dispatchers with the key information for pre-DR resource assignment of the next day. We evaluate the MAPE of 24-step forecasting on 300 buildings, as shown in Fig. 6. Since the consumption in the future is more difficult to forecast than that at just next moment, the multi-step forecasting is more error-prone than the single-step one. The proposed model has an average MAPE reduction of 31.2%, 30.5%, 37.3%, and 22.7% relative to GRU, ResNet, LSTM, and GCNN, respectively.

C. Probabilistic Forecasting

Probabilistic forecasting can provide more information than point forecasting, thus providing more application possibilities. Common probability forecasting include probability distribution parameter estimation, probability interval estimation, and quantile forecasting.

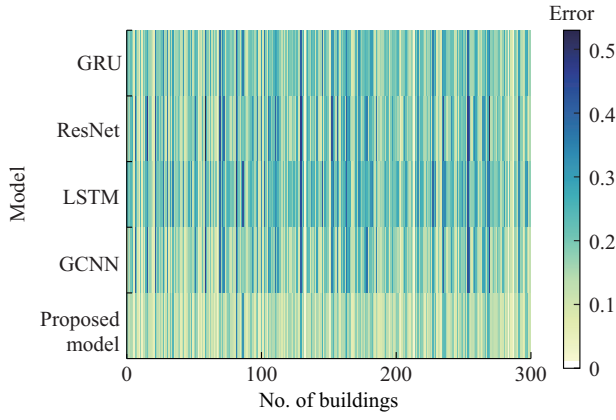


Fig. 6. MAPE heat map of 24-step forecasting on 300 buildings.

Among them, quantile forecasting does not depend on the distribution hypothesis, and the probability interval can be generated according to quantile forecasting. Therefore, quantile forecasting has a better generalization ability and forecasting accuracy. In this study, we use pinball loss as the loss function of each model and generate the forecasting of 9 quantiles from 0.1 to 0.9. By comparing the average pinball scores, we can observe the accuracy improvement brought by the proposed model for probabilistic forecasting.

Pinball scores of several models are presented in Fig. 7. The comparison on the three buildings shows that the proposed model has a significant reduction in pinball score compared with other models, and the improvement is greater than the MAE improvement of point forecasting. Compared with the MSE loss function, pinball loss holds a more complex search space, making the model more difficult for conventional deep models to converge on an acceptable local optimal point. The experimental results show that although our network structure is more complicated than ordinary CNN, it still achieves a better performance in probability forecasting. In view of the importance of probabilistic forecasting in current load forecasting, our proposed model presents a promis-

ing future.

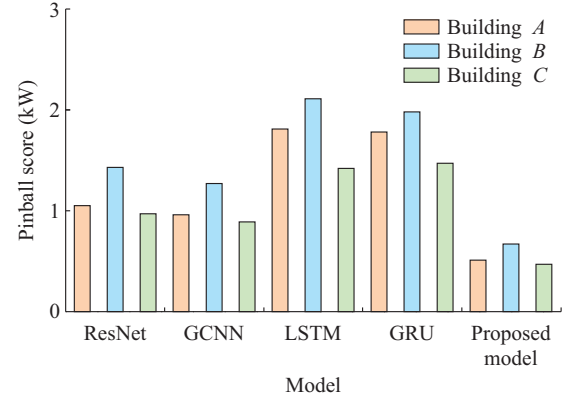


Fig. 7. Pinball scores of probabilistic load forecasting for three buildings.

D. Fusion Method

The TFF network also contains two branches, one for learning historical load sequences and the other for learning external factors. The difference with our model is that the representative vectors output from the two branches are only fused in the manner of vector concatenating, thus a fully-connected network is also needed in the rear for learning the vector obtained by concatenating. In this subsection, we compare our fusion model with TFF, keeping the other structures and parameters of two models consistent.

The comparison of the results is shown in Table VII. The RMSE of the TFF network is superior to that of our model for the three buildings, while its MAPE and MAE are also superior to other models on buildings *A* and *C*. It proves that the TFF network is effective. However, the MAPE, MAE, and RMSE of our proposed model have better performances than the those TFF network in the three buildings, which reveals the relationship between sequence features and external factor features is modeled effectively. The pinball score of the probability forecasting also shows the similar conclusion.

TABLE VII
COMPARISON OF PROPOSED MODEL AND TFF

Building	MAPE (%)		MAE (kW)		RMSE (kW)		Pinball score (kW)	
	TFF	Proposed model	TFF	Proposed model	TFF	Proposed model	TFF	Proposed model
<i>A</i>	2.51	2.22	1.85	1.43	2.46	2.09	0.92	0.51
<i>B</i>	5.36	2.80	2.94	1.63	3.42	2.27	1.04	0.67
<i>C</i>	6.52	4.26	2.48	1.25	3.17	1.67	0.78	0.47

E. Comparison of Computation Efficiency

Table VIII summarizes the computation time of all models for three buildings. The computation efficiency of our proposed model is the fastest. LSTM and GRU consume a lot of time due to the characteristics of their recursive calculations. Compared with ResNet and GCNN which have similar structures to ours, our proposed model converges faster and achieve a better accuracy.

TABLE VIII
COMPARISON OF COMPUTATION EFFICIENCY WITH FIVE MODELS

Building	Computation time (s)				
	GRU	ResNet	LSTM	GCNN	Proposed model
<i>A</i>	712	348	828	528	280
<i>B</i>	756	312	864	576	204
<i>C</i>	782	420	880	672	286

This is because our network explicitly models the relationship between the temperature and load curve, making it easier to converge to a more optimal extreme value. Figure 8 compares the learning curves of the models, using the dataset of building *A*. It reveals our proposed model has the lowest training error and the fastest rate of convergence.

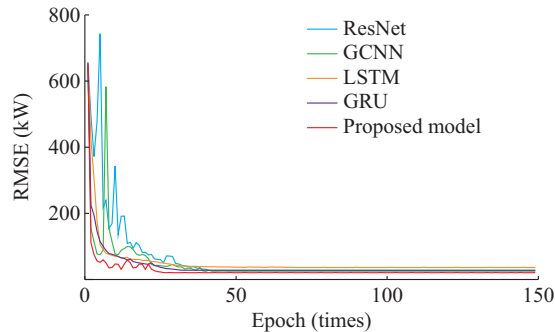


Fig. 8. Learning curves of training error for building *A* with five models.

V. CONCLUSION

In this paper, we propose a DCNN based on ResNet for building load forecasting. With the dilated convolution in forecasting branch, the ability of CNN has been improved by extracting complex and significant features of load sequences. In addition, we propose another external factor integration branch which takes more significant weather features as the input. The features extracted from external factors will be fused effectively to enhance the ability of learning discriminative features remarkably. Therefore, the forecasting accuracy is optimized greatly without increasing parameters and operations. In this study, the performances of five different deep learning models, i.e., GRU, ResNet, LSTM, GCNN and our proposed model in the application of single-step and 24-step building load forecasting are systematically compared. Competitive results reveal that our model can serve more accurate forecasting, higher computational efficiency, and stronger generalization for different buildings.

REFERENCES

- [1] U.S. Energy Information Administration. (2020, Nov.). U.S. energy information administration monthly energy review. [Online]. Available: <http://www.eia.gov/totalenergy/data/monthly/#consumption>
- [2] K. Amber, R. Ahmad, M. Aslam *et al.*, "Intelligent techniques for forecasting electricity consumption of buildings," *Energy*, vol. 157, pp. 886-893, May 2018.
- [3] F. Wang, H. Xu, T. Xu *et al.*, "The values of market-based demand response on improving power system reliability under extreme circumstances," *Applied Energy*, vol. 193, pp. 220-231, Jan. 2017.
- [4] I. Atzeni, L. G. Ordóñez, G. Scutari *et al.*, "Demand-side management via distributed energy generation and storage optimization," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 866-876, Jun. 2013.
- [5] Q. Shi, F. Li, Q. Hu *et al.*, "Dynamic demand control for system frequency regulation: concept review, algorithm comparison, and future vision," *Electric Power Systems Research*, vol. 154, pp. 75-87, Jan. 2018.
- [6] F. Kienzie, P. Ahcin, G. Andersson *et al.*, "Valuing investments in multi-energy conversion, storage, and demand-side management systems under uncertainty," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 2, pp. 194-202, Apr. 2011.
- [7] R. Jain, K. Smith, P. Culligan *et al.*, "Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Applied Energy*, vol. 123, pp. 168-178, Jun. 2014.
- [8] A. Colmenar-Santos, L. N. T. de Lober, D. Borge-Diez *et al.*, "Solutions to reduce energy consumption in the management of large buildings," *Energy & Buildings*, vol. 56, pp. 66-77, Jan. 2013.
- [9] N. Somu, G. Rama, and K. Ramamritham, "A hybrid model for building energy consumption forecasting using long short term memory networks," *Applied Energy*, vol. 261, p. 114131, Mar. 2020.
- [10] H. Zhao and F. Magoul's, "A review on the prediction of building energy consumption," *Renewable & Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586-3592, Aug. 2012.
- [11] K. Song, Y.-S. Baek, D. Hong *et al.*, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Transactions on Power Systems*, vol. 20, no. 1, pp. 96-101, Feb. 2005.
- [12] S. M. Moghaddastafreshi and M. Farhadi, "A linear regression-based study for temperature sensitivity analysis of Iran electrical load," in *Proceedings of IEEE International Conference on Industrial Technology*, Chengdu, China, Apr. 2008, pp. 1-7.
- [13] H. Tao, G. Min, M. E. Baran *et al.*, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," in *Proceedings of IEEE PES General Meeting*, Providence, USA, Jul. 2010, pp. 1-8.
- [14] L. Wei and Z. G. Zhang, "Based on time sequence of ARIMA model in the application of short-term electricity load forecasting," in *Proceedings of 2009 International Conference on Research Challenges in Computer Science*, Shanghai, China, Dec. 2009, pp. 11-14.
- [15] C. M. Lee and C. N. Ko, "Short-term load forecasting using lifting scheme and ARIMA models," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5902-5911, May 2011.
- [16] W. Hong, "Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm," *Energy*, vol. 36, no. 9, pp. 5568-5578, Sept. 2011.
- [17] W. Li, X. Yang, H. Li *et al.*, "Hybrid forecasting approach based on GRNN neural network and SVR machine for electricity demand forecasting," *Energies*, vol. 10, no. 1, p. 44, Jan. 2017.
- [18] D. Basak, P. Srimanta, and D. C. Patranbis, "Support vector regression," *Neural Information Processing Letters and Reviews*, vol. 11, no. 10, pp. 203-224, Sept. 2007.
- [19] F. Cheng, X. Fu, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Applied Energy*, vol. 127, pp. 1-10, Aug. 2014.
- [20] F. Zhang, C. Deb, S. E. Lee *et al.*, "Time series forecasting for building energy consumption using weighted support vector regression with differential evolution optimization technique," *Energy & Buildings*, vol. 126, pp. 94-103, Aug. 2016.
- [21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
- [22] T. Abel, P. V. Nguyen, M. Barad *et al.*, "Genetic demonstration of a role for PKA in the late phase of LTP and in hippocampus-based long-term memory," *Cell*, vol. 88, no. 5, pp. 615-626, Mar. 1997.
- [23] W. Kong, Z. Y. Dong, Y. Jia *et al.*, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.
- [24] W. Kong, Z. Y. Dong, D. J. Hill *et al.*, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087-1088, Mar. 2017.
- [25] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting novel pooling deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271-5280, Sept. 2017.
- [26] Z. Deng, B. Wang, Y. Xu *et al.*, "Multi-scale convolutional neural network with time-cognition for multi-step short-term load forecasting," *IEEE Access*, vol. 7, pp. 88058-88067, Jul. 2019.
- [27] Z. Deng, B. Wang, H. Guo *et al.*, "Unified quantile regression deep neural network with time-cognition for probabilistic residential load forecasting," *Complexity*, vol. 2020, pp. 1-18, Jan. 2020.
- [28] B. Xu, N. Wang, T. Chen *et al.* (2015, Nov.). Empirical evaluation of rectified activations in convolutional network. [Online]. Available: <https://arxiv.org/abs/1505.00853>
- [29] S. Ioffe and C. Szegedy. (2015, Mar.). Batch normalization: accelerating deep network training by reducing internal covariate shift. [Online]. Available: <https://arxiv.org/abs/1502.03167>
- [30] C. Miller and F. Meggers, "The building data genome project: an open, public data set from non-residential building electrical meters," *Energy Procedia*, vol. 122, pp. 439-444, Sept. 2017.
- [31] H. Wang, Y. Wang, Q. Zhang *et al.*, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sensing*, vol. 9, no. 5, p. 446, May 2017.

Jinsong Wang received the B.S. degree in computer science from Kent State University, Kent, USA, in 2015. He is currently pursuing the Ph.D. degree in system and control engineering at Case Western Reserve University, Cleveland, USA. His research interests include artificial intelligence, load forecasting and fault diagnosis.

Xuhui Chen received the B.E. degree in information engineering from Xidian University, Xi'an, China, in 2012, the M.S. degree in electrical and computer engineering from Mississippi State University, Starkville, USA, in 2015, and the Ph.D. degree in computer engineering from Case Western Reserve University, Cleveland, USA, in 2019. She joined the College of Aeronautics and Engineering as an Assistant Professor at Kent State University, Kent, USA, in 2019. Her research interests include cyber security and artificial intelligence.

Fan Zhang received the B.S. degree in power systems from Hunan Univer-

sity, Changsha, China, in 2013, and the M.S. degree from Case Western Reserve University, Cleveland, USA, in 2016. She is currently pursuing the Ph.D. degree in the Department of Electrical Engineering and Computer Science, Case Western Reserve University. Her research interests include transmission system planning, larger-scale optimization algorithms and smart grid.

Fangxi Chen received the B.S. degree from the Shenyang University of Technology, Shenyang, China, in 2017. She is currently pursuing the M.S. degree in Northeastern University, Shenyang, China. Her research area is artificial intelligence in load forecasting.

Yi Xin is studying in the major of software engineering, in College of Software, Northeastern University, Shenyang, China. His main research interests include load forecasting, computer vision and data mining.