IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 October 2022, accepted 26 October 2022, date of publication 3 November 2022, date of current version 8 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3218374

RESEARCH ARTICLE

Short-Term Load Forecasting Based on Improved TCN and DenseNet

MINGPING LIU[1], HAO QIN[1], RAN CAO[1], AND SUHUI DENG[1,2]
[1]School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]Jiangxi Provincial Key Laboratory of Interdisciplinary Science, Nanchang University, Nanchang 330031, China

Corresponding author: Suhui Deng (shdeng@ncu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61865011 and Grant 62065012; in part by the Natural Science Foundation of Jiangxi Province of China under Grant 20212BAB202031; in part by the Interdisciplinary Innovation Fund of Natural Science, Nanchang University, under Grant 9167-28220007-YB2111; and in part by the Innovation Fund Designated for Graduate Students of Jiangxi Province of China under Grant YC2021-S152.

ABSTRACT With the grid-connected application of renewable energy sources such as wind and photovoltaic power, the nonlinearity and fluctuation of load data makes load forecasting more difficult than ever before. In order to extract the implicit relationship between multiple features and power load to construct a long-term sequence dependency, this paper proposes a short-term load forecasting based on improved temporal convolutional network (TCN) and densely connected convolutional network (DenseNet). Firstly, multiple features are reconstructed by using a fixed-length sliding window, and then the high-dimensional features reflecting the complex and non-stationary characteristics of power load are extracted by the DenseNet to construct a feature matrix. Secondly, we innovatively improve the TCN and introduce a parallel pooling into the traditional TCN to mine the features of time sequences. Finally, the self-attention mechanism (SAM) is used to further enhance the weight of key features to eliminate the influences of interference signals. Experiments were performed on Southern China and ISO-NE (New England) public datasets to verify the effectiveness and generalization of the proposed model. Compared with the traditional TCN, the mean average percentage error (MAPE) of the improved TCN on the two datasets decreases by 23.38% and 8.14%, respectively. Furthermore, when compared to the TCN-SAM hybrid model, the MAPE of the proposed model is significantly reduced by 42.41% and 26.89%, respectively.

INDEX TERMS Short-term load forecasting, improved temporal convolutional network, densely connected convolutional network, self-attention mechanism.

I. INTRODUCTION

High accuracy of load forecasting is essential for the generation, transmission, distribution and consumption of electric energy. However, the intermittence of renewable energy sources and the randomness of electric vehicles inevitably increase the complexity and uncertainty of the power systems. Therefore, it is still a challenging task to obtain high accuracy of short-term load forecasting (STLF). Furthermore, STLF plays a vital role in balancing power system supply and load demand to avoid the instability of power grid [1]. Accurate load forecasting can not only control the consumption behavior of power demand-sides in real time to ensure the reliability of power supply and improve economic benefits, but also help to provide scientific guidance for day-to-day operation of power systems. Meanwhile, recent researches showed that increasing the load forecasting error by 1% raised millions of dollars for the power industry [2]. The fluctuation of short-term power load sequence has obvious randomness and nonlinearity, and the influencing factors, such as temperature, electricity price, and holidays, are diversified and complex. All these will bring huge challenges to accurate forecasting [3].

There has been much in-depth research on the methods to obtain higher accuracy and generalization of STLF. These methods are divided into traditional statistical models

The associate editor coordinating the review of this manuscript and approving it for publication was Tony Thomas.

and machine learning models in general [4]. The traditional statistical methods have been widely used to forecast power load in the early days. It mainly includes exponential smoothing (ES) [5], [6], autoregressive integrated moving average (ARIMA) [7], [8], grey model (GM) [9], Kalman filter (KF) [10], [11], etc. Although these models have been useful in dealing with linear forecasting relationships, they were not effective ways to accurately predict the nonlinear time series with huge datasets. The accuracy of load forecasting based on statistical methods drops when a longer length of the prediction time horizon is needed [12]. With the development of computer technology, a number of machine learning models have been successfully proposed to load forecasting by extracting nonlinear features. Fan et al. [13] proposed an adaptive method of support vector machine (SVM). The model classified the input data into several subsets in an unsupervised way and fitted the input data to different market states in a supervised way by using the SVM. Liu et al. [14] proposed a hybrid STLF model based on improved fuzzy C-means clustering, random forest and deep neural networks to significantly improve the prediction performance of holidays. Cecatiet al. [15] used a novel error correction algorithm to train radial basis function (RBF) that weakened the interactions between RBF units and then reduced the input pattern on RBF. The experimental results showed that the proposed model had superior performance compared to other state-of-the-art machine learning methods. A constrained quantile regression average (CQRA) method was proposed in [16], which could create an improved integration from several independent probability predictions. In [17], an unspecified nonlinear relationship between load and weather variables was established to STLF by using artificial neural network (ANN). However, because these methods mentioned above are still difficult to extract the features of time series, we cannot generalize these methods to different kinds of datasets mainly due to the small number of parameters [18], [19].

In recent years, deep learning technologies have been popularly used in the forecasting of time series due to their ability of extracting in-depth features in huge datasets by multilayer nonlinear mapping during training stage. Then, these methods can better fit the nonlinear relationships between input and output to enhance the superior performance of load forecasting [20]. Recurrent neural network (RNN) is one of the commonly used deep learning models for the forecasting of time series [21]. Furthermore, its variants, e.g., long short-term memory network (LSTM) and gated recurrent unit network (GRU), have better performance than the traditional RNN and other load forecasting techniques. Kong et al. [22] applied the LSTM to predict the power load of a single household. The experimental results showed that the prediction accuracy was better than those of other models. To improve the accuracy of prediction, Wu et al. [23] used the GRU to forecast short-term load considering the impact of electricity price. Kong et al. [24] utilized deep belief network (DBN) and genetic algorithms to optimize the parameters of network. Although these models for the forecasting of time series have been successfully applied to load forecasting, there are still some shortcomings in data processing, extracting in-depth features and dealing with long-time series. In terms of these, hybrid models are being increasingly used to enhance the accuracy of load forecasting. In [25], a novel ResNet improved with probability prediction, Monte Carlo dropout and SELU activation function was adopted to efficiently extract deep features. The validity and generalization of the hybrid model were successfully verified on three public datasets. Tang et al. [26] proposed a short-term load forecasting model based on temporal convolutional network (TCN) with channel and temporal attention mechanism (AM), which fully exploited the nonlinear relationship between meteorological factors and load. The maximum information coefficient (MIC) was adopted to select the high-quality variables of input and eliminate irrelevant variables to reduce the parameters. In [27],the hidden information in the features was extracted by the TCN and the relationship of time series was constructed. Moreover, the load of industrial users was predicted by using the LightGBM.

Nevertheless, although these models mentioned above have achieved good performance, it struggles to simultaneously extract both the features of time series and deep features. In the aspect of extracting the features of time series, the LSTM has shown unprecedented advantages in sequence modeling tasks. However, the LSTM has some inherent drawbacks, including the vanishing or exploding of the gradients and the inability to process in parallel [28]. The TCN is a new CNN-based model for analyzing sequences that mainly contains three modules, i.e., causal convolution, dilated convolution and residual block [29]. Moreover, the TCN can be capable in capturing long-range dependencies between the load series by relying on a large receptive field, and its residual blocks help to avoid gradient explosion [30]. However, the TCN cannot extract the internal correlation information of input [31] and the loss of local information caused by dilated causal convolution, which results in preventing to further improve the accuracy of STLF [32].

To overcome these disadvantages, we propose a novel method called the TCN-DenseNet based Network. Firstly, the DenseNet is used to extract the internal correlation information of input. It can avoid the vanishing of gradient and strengthen the transfer of features. The number of parameters are substantially reduced to improve the training efficient [33]. Secondly, the parallel pooling structure is introduced into the residual block of the TCN to retain the feature map needed for prediction through the translation and rotation invariance, so as to improve the robustness of the model and reduce the loss of information. In other words, the proposed model uses the DenseNet to extract the hidden features of input and the improved TCN (iTCN) is adopted to extract the time features of long-term load series. Finally, the SAM is used to enhance these different features. The experimental results performed on two open datasets show that the hybrid model based on the DenseNet-iTCN has higher

accuracy and generalization than other prevalent methods in the area. The main contributions of this paper are described as follows:

1) We analyze the nonlinear correlation between various meteorological features and load series. A fixed-length sliding time window is used to capture the actual variation and fluctuation trend in the features.
2) We introduce a parallel pooling structure to improve the residual module of the TCN. The iTCN reduces the loss of information and performs better in extracting long-term temporal relationships. The DenseNet is applied to extract internal correlation information of input and other related factors such as temperature, holiday, and day types.
3) The SAM is adopted in the proposed model to mine the relationship between load series and input features, so that the proposed model can focus on the key information to achieve better performance and realize the high accuracy of load forecasting.
4) A load forecasting model based on the DenseNet-iTCN is proposed. This paper extends the application of the proposed model to two datasets from Southern China and New England, USA. Experimental results demonstrate that the proposed model achieves better performance than other existing models in STLF.

The rest of this paper is organized as follows. Section II presents the basic theory of each algorithm and the structure of the DenseNet-iTCN framework. In Section III, we introduce the experimental setting and compare the proposed model with other existing models on two public datasets. The conclusion of this paper and the future work are drawn in Section IV.

II. METHODOLOGY

In this section, we propose an ensemble framework for STLF that consists of three main modules, i.e., DenseNet, improved TCN and self-attention mechanism. The following subsections briefly describe each module used in this study. Finally, we will give the architecture of the proposed hybrid model, in which the feature selection, data reprocessing, and the evaluation criteria will be described.

A. DENSENET

With the deepening of the CNN, there will be some problems to train the model, such as the vanishing or exploding of gradients, network degradation, etc. The ResNet [34] can overcome these problems mentioned above through the residual connection. However, the ResNet combines features of all preceding layers through summation before they are passed into a current layer. Moreover, the number parameters of ResNet is substantially larger because each layer has its own weights. The DenseNet retains the idea of the construction of residual module. Each layer obtains additional inputs from all preceding layers and delivers its output of features to all subsequent layers, which improves the information transmission capacity of each layer [35]. In this case, the convolution network with L layers has $L(L+1)/2$ connections, instead of only L connections for the traditional models. Therefore, the DenseNet has significantly improved compared to other CNN-based models. It has been widely used in medical treatment [36], [37], semantic segmentation [38], image recognition and classification [39], [40], audio processing [41], and other fields.

The dense block is the basic unit of the DenseNet, and its structure is schematically shown in Figure 1. Accordingly, the feature maps of all preceding layers at the l^{th} layer can be given as

$$x_l = H_l\left([x_0, x_1, \cdots, x_{l-1}]\right) \tag{1}$$

where x_l represents the feature map generated in the l^{th} layer and $[x_0, x_1, \cdots, x_{l-1}]$ indicate the concatenation of all preceding layers. $H_l(\cdot)$ represents batch normalization (BN) followed by a rectified linear activation function (ReLU) and a 3×3 convolution.

The transition block connects two adjacent dense blocks, which consists of a 1×1 convolution layer and a 2×2 average pooling layer. The transition block reduces the size of the feature map by down sampling and compresses the model so as to reduce the amount of computation and improve the computational efficiency.

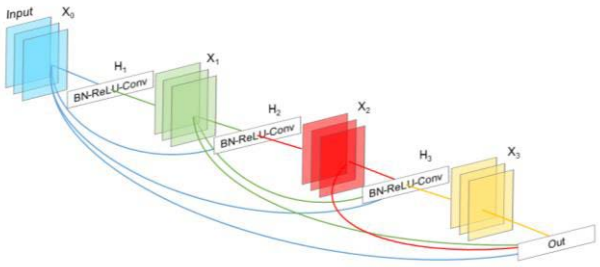

FIGURE 1. The schematic structure of the dense block.

B. IMPROVEDTCN

The size of convolution kernel limits the ability of traditional convolution network to extract the features of long-term series of load data. In order to conquer the problems mentioned above, the TCN is introduced due to its innovative integration of causal convolution, dilated convolution and residual block. Figure 2 shows the overall flow chart and architecture of the TCN model, each module of which will be described in the following.

1) CAUSAL CONVOLUTION

The TCN developed from CNN-based model overcomes the disadvantages of LSTM in solving time series. The output sequence of TCN has the same length as the input sequence and it uses the information from the preceding time steps. That is to say, the output of the network is only related to the previous input, avoiding the disclosure of future information. Taking the structure of causal convolutions shown in Figure 3 as an example, the output y_t at time step t depends

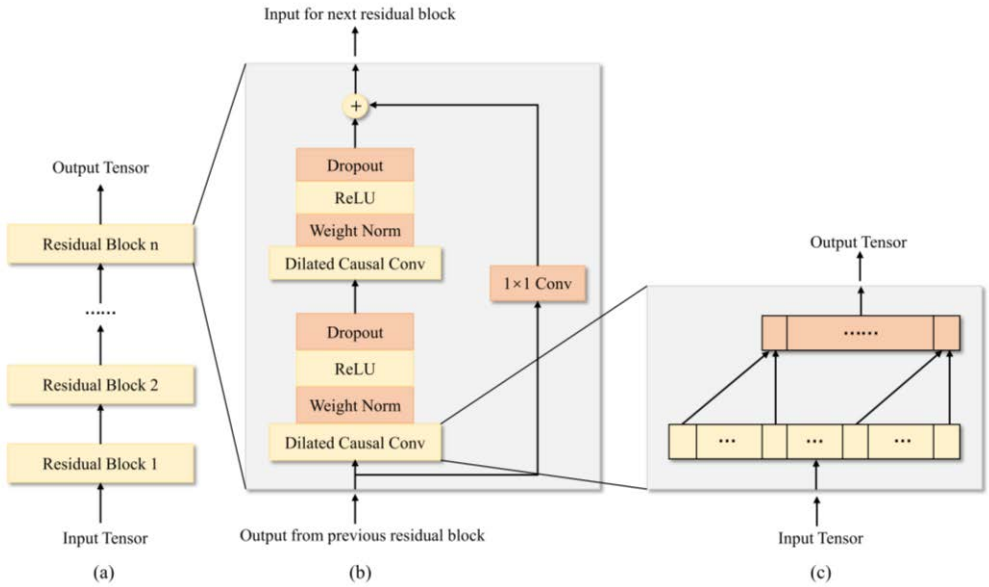

FIGURE 2. (a) Deep TCN. (b) Residual block. (c) Brief diagram of the dilated convolution.

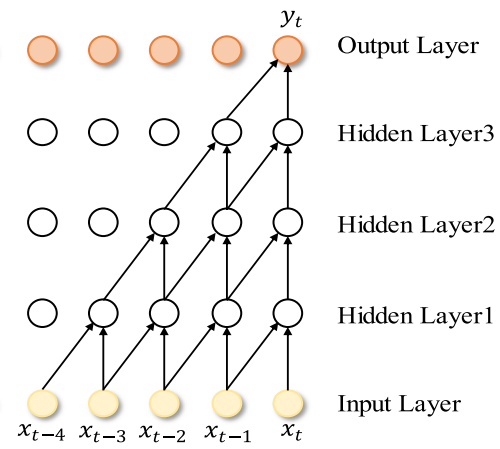

FIGURE 3. The structure of the causal convolution.

on the previous input $(x_{t-4}, x_{t-3}, x_{t-2}, x_{t-1}, x_t)$ but has nothing to do with future input $(x_{t+1}, x_{t+2}, \cdots, x_T)$. The causal convolution utilizes a one-dimensional full-convolutional network to keep the input and output data of the same length by adding zero padding.

2) DILATED CONVOLUTION

Although the receptive field of the network can be expanded by stacking the causal convolutions, the number of causal convolution layers should be large enough to train the long time series of input data. In this case, it will result in the vanishing of gradients and low computational efficiency. In order to overcome these disadvantages, the dilated convolutions have been introduced into the TCN to increase exponentially the receptive field of the network. By sampling the upper input at internal and increasing exponentially the dilated factor with $d = 2^i$(i is the number of the network layers), the TCN can achieve as large receptive field as possible with fewer layers of the network. The dilated convolution is defined as follows:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i} \tag{2}$$

where d is the dilated factor, k is the convolution kernel size, and $x_{s-d \cdot i}$ means only the convolution of the past state. An illustration of the dilated convolution with kernel size $k = 2$ and dilated factor $d = [1, 2, 4, 8]$ is shown in Figure 4. The dilated convolution with 5 layers can read 16 inputs whereas an ordinary network with 16 layers will be used to obtain the same receptive field. Therefore, the dilated convolution increases the receptive field of the network without increasing the parameters, which reduces the network complexity and improves the computational efficiency.

3) RESIDUAL BLOCK

With the deepening of the network, the parameters of the network model increase, which results in the vanishing or explosion of gradients. Thus, the residual connection has been added into the TCN to ensure its stability as the increase of network layers. The structure of residual module is shown in Figure 2(b). One can find that there are two layers of the dilated causal convolution and ReLU in each residual block. Furthermore, the weight normalization layer and dropout layer after dilated causal convolution are used to improve the generalization of the network. Finally, a 1×1 convolution is adopted to ensure the same dimension of the input and output. Figure 2(a) shows a deep TCN formed by stacking n residual blocks, which can extract the features of long-term historical series. That is to say, each convolution of the output layer

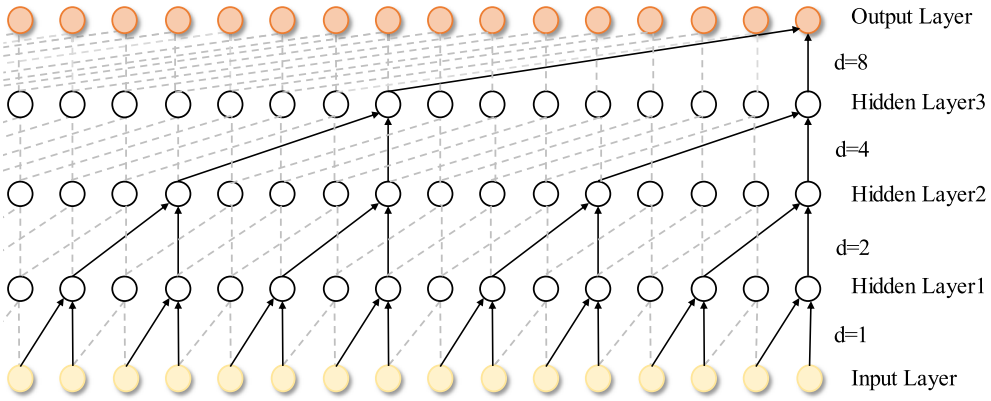

FIGURE 4. The structure of the dilated convolution.

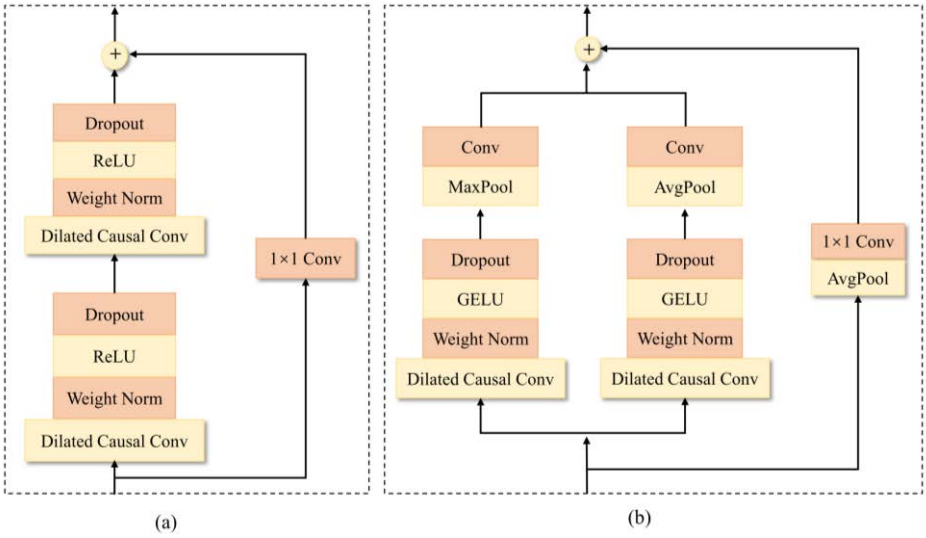

FIGURE 5. (a) The structure of the original TCN. (b) The structure of the iTCN.

receives more information from the convolution of the input layer.

4) ITCN

At present, many researchers have also proposed meaningful improvements to TCN. In [42], the original TCN was modified to a parallel structure with two branches for mechanical fault diagnosis. Li et al. [43] proposed a novel MS-TCN++ model with a multi-stage architecture to capture temporal dependencies and reduce over-segmentation errors, in which the first one was the prediction generation stage and other ones were the refinement stages. Korkmaz et al. [44] designed a novel CNN with a parallel pooling structure, which consisted of max-pooling and average-pooling blocks, to increase the performance of the forecasting. In this paper, we add the parallel pooling structure in the residual module of the TCN under the motivation coming from [43] and [44]. In addition, the ReLU activation function was replaced with GELU [45]. As a comparison, Figure 5 shows the original TCN and the improved TCN (iTCN) with the parallel pooling structure.

C. SELF-ATTENTION MECHANISM

Attention mechanism is generated by simulating human visual attention. Human inevitably pay attention to several key parts when observing things. Therefore, the attention mechanism is designed based on this phenomenon to learn the important features of key parts and then splices them together. Furthermore, feature enhancement is carried out by weighted summation. The SAM [46], i.e., a variant of the attention mechanism, focuses on the relevance of internal features and gives different weights according to the importance of input features. The input sequence $X = [x_1, x_2, \cdots, x_T]$ is linearly transformed with three different weight matrices W_q, W_k, and W_v to get query (Q), key (K), and value (V). The similarity is calculated between Q and K, and then the results are

normalized by a softmax function to obtain the self-attention matrix (W). Finally, one can multiply the obtained self-attention matrix by the matrix V to get the output matrix [47]. The SAM is defined as follows:

$$\begin{cases} Q = W_q X \\ K = W_k X \\ V = W_v X \end{cases} \tag{3}$$

$$W = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4}$$

$$\text{Attention}\,(Q, K, V) = WV = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where d_k is the dimension of K, $(d_k)^{-1/2}$ is the scaling factor, and softmax$(\cdot)$ is the function of normalization by column.

D. LOAD FORECASTING FRAMEWORK BASEDON DENSENET-ITCN

The deep structure of the DenseNet could extract the complex relationships of power load with time, temperature, humidity, and other characteristics. Due to the integration of the extraction ability of the CNN and the time-domain modeling ability of the RNN, the TCN can extract the time correlation in features [48]. At the same time, the self-attention mechanism can enhance the characteristics extracted by the DenseNet and iTCN. In order to better map the relationship between input and output, the overall framework of the DenseNet-iTCN is shown in Figure 6, which consists of data preprocessing, feature selection, and evaluation methods. The process of the proposed model can be concisely explained in Algorithm 1. Firstly, the raw data is divided into the training set, validation set and testing set at a rate of 8:1:1. Secondly, the parameters of DenseNet, iTCN and SAM algorithms are initiated. Thirdly, if the validation loss of the prediction model decreases after training in each epoch, it should adjust all the parameters until the epochs are incremented. However, if the validation loss does not decrease for 30 epochs, the training process then will stop. Finally, the optimal model will be evaluated on the testing data.

1) FEATURE SELECTION

There are a multitude of external factors that influence the prediction accuracy of power load. These external factors are usually complex and diverse, which lead to the dynamic or random trend of load series. For example, the government policies may lead to the fluctuation of power load and temperature also gives rise to the fluctuation in power consumption. At the same time, the changing trend of power load has a certain regularity and periodicity due to the regular social activities and industrial production. The demand change of the dataset for a month in a region of Southern China is shown in Figure 7. It is clear that the trend of the power load has a 7-day period, during which the power load from Monday to Friday fluctuates slightly and is higher than that on weekend.

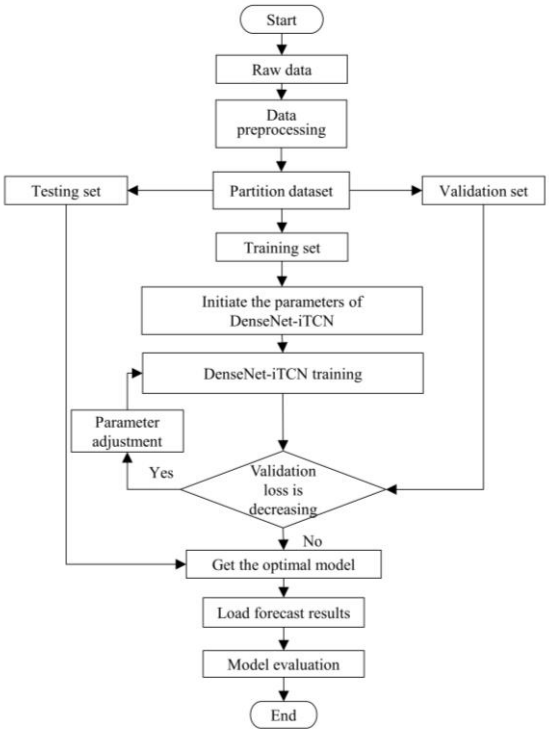

FIGURE 6. The overall framework of the proposed model.

Algorithm 1 Algorithm for the DenseNet-iTCN

Input: Raw data [D, T, H, W, S].
Output: Trained model G.
1: The raw data is divided into training set, validation set and testing set.
2: Initiate the parameters of DenseNet, iTCN and self attention.
3: For i =1 to N do
$x \leftarrow H_l\left(\left[x_0, x_1, \cdots, x_{l-1}\right]\right)$
$f(s) \leftarrow (x *_d f)(s)$ #dilated convolution
$y(x) \leftarrow f(s) + x$ #residual
$d = \text{softmax}(y(x))$
compute *loss*
$P \leftarrow \text{Adam}(P, r)$ # r=learning rate
End for
4: If v_1 stop decreasing or N > maximum iterations
$\#v_1$ =validation loss
End if
5: Return trained model G

Furthermore, the day types, e.g., weekday, holiday and special festival, also have significant impact on the consumption of power load. Figure 8 shows the profiles of the power load and temperature in ISO-NE for 6 years. One can find that the trend of power load has 12 peaks and valleys with the period of six months, because the power consumption reaches peak as the temperature is very high or low. Therefore, the temperature feature is an important characteristic of load forecasting in the proposed model. Furthermore, figure 9 shows the correlation matrix among the load data and external factors. It is obviously that weekday, hour, temperature and demand

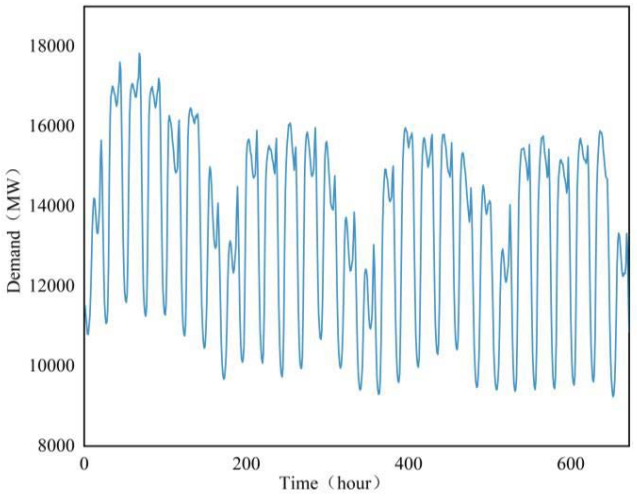

FIGURE 7. Daily load profile of 4 weeks.

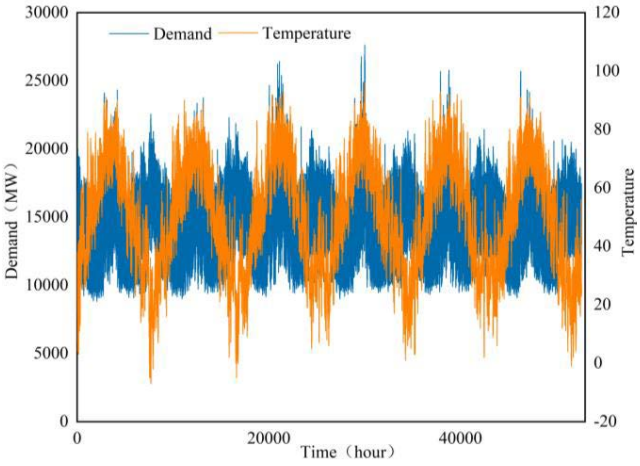

FIGURE 8. Power load and temperature profiles of 6 years.

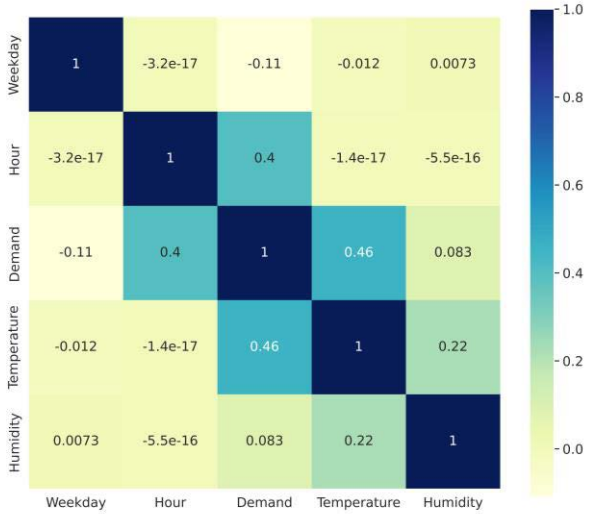

FIGURE 9. Correlation matrix among the power load and external factors.

TABLE 1. Features for the load forecast.

Symbol	Feature	Description of the feature	
D	Demand	Normalize the demand of power load	
T	Temperature	Normalize the temperature	
H	Holiday	Yes	[1,0]
		No	[0,1]
W	Weekend	Yes	[1,0]
		No	[0,1]
S	Season	Spring	[1,0,0,0]
		Summer	[0,1,0,0]
		Autumn	[0,0,1,0]
		Winter	[0,0,0,1]

are moderate correlation, and their correlation coefficients are 0.4 and 0.46, respectively.

2) DATA PREPROCESSING

In view of the autocorrelation, periodicity, and trend of power demand data, this paper chooses the five characteristics of power load, temperature, holiday, season, and weekend for load forecasting. Each type of features is processed as shown in Table 1 and the one-hot encoder is adopted for season, holiday, weekend, and weekday. However, the demand of power load and temperature will be normalized by min-max to scale the data to range [0 1]. The min-max normalization can be defined as

$$\hat{x} = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{6}$$

where X is the whole time series, x is the data before normalization, and $\hat{x}$ is the data after normalization.

In this paper, a fixed-length sliding window is used to extract data as the input of the network. It takes one hour as the moving step. The structure of sliding window is shown in Figure10. It is well known that the characteristics of adjacent time points have a great influence on forecasting the power load of the next time point. In order to determine the appropriate time range, we used the proposed model to predict the load demand of the next hour through the features of the first 12, 24, and 48 hours, respectively. From the results of the three models as shown in Table 2, it is clear that the characteristics of the first 24 hours are appropriate, and the experimental results are better than the others.

The sliding window has a length of 24 hours and a step size of 1 hour. In Figure 10, D_1 represents a collection of demand in the first 24 hours, i.e., $[d_1, d_2, \cdots d_{24}]$, D_2 is a collection of demand from the 2nd hour to 25th hour, i.e., $[d_2, d_3, \cdots d_{25}]$, and D_t means $[d_t, d_{t+1}, \cdots d_{t+23}]$. In addition, the parameters T, H, S, and W present a 24-hour feature matrix of temperature, holiday, season, and weekend, respectively.

The data split by the sliding window are successively fed into the DenseNet and then its output is further enhanced by the SAM. The enhanced feature matrix is input into the iTCN to construct the timing relationship for extracting the features

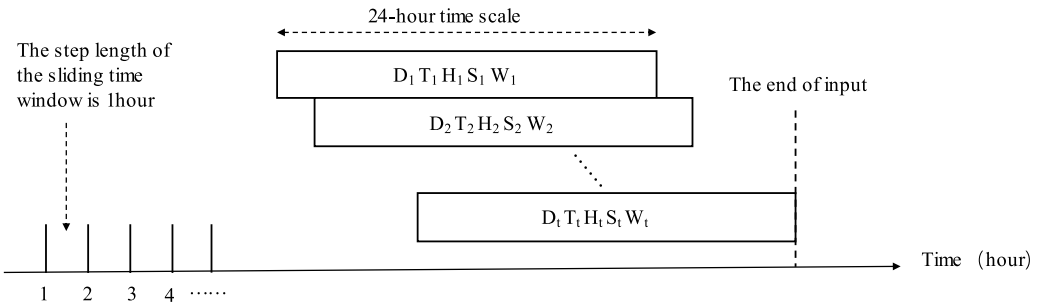

FIGURE 10. Fixed-length sliding window.

TABLE 2. The statistical metrics of the three models for different time horizon.

	12hours			24hours			48hours		
	MAPE(%)	MAE(MW)	RMSE(MW)	MAPE(%)	MAE(MW)	RMSE(MW)	MAPE(%)	MAE(MW)	RMSE(MW)
TCN	1.68	245.17	359.49	1.26	187.81	303.22	1.29	187.38	298.56
LSTM	1.50	228.06	344.51	1.37	203.49	303.78	1.43	212.77	313.85
Ours	1.01	149.24	207.07	0.86	126.99	171.19	0.99	145.47	197.08

of time series, and then the output of iTCN is entered into the SAM again. Finally, the prediction results can be achieved from the full connection layer.

3) PERFORMANCE EVALUATION

In order to evaluate the performance of the proposed model, the mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE) are used as evaluation indices. The statistical metrics are defined as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \tag{7}$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{9}$$

where y_t and $\hat{y}_t$ represent the real value and the predicted value, respectively. The MAPE is used to prove the accuracy of the model. The smaller the value of MAPE is, the higher the accuracy of the proposed model is. The MAE shows the robustness of the proposed model to outliers. Therefore, it is necessary to adopt these statistical metrics to evaluate the performance of the proposed model.

III. EXPERIMENT AND RESULT ANALYSIS

A. DESCRIPTION OF DATASET

The first data were derived from the state grid of a region in Southern China with sampling every 15 minutes [49], i.e., sampling 96 points a day. There are 55000 sets of data from January 1, 2012 to July 6, 2013.The datasets are split into a training set, validation set and testing set according to the proportion of 8:1:1. A total of 42848 sets are used for training the model, 7152 sets can be validated the model and 5000 sets are defined as the testing set.

The second public dataset were collected from ISO-NE (New England) dataset [50], which included New England's electric load from March 2003 to December 2014 with one-hour resolution. The dataset contain the power load and temperature data. A total of 35232 sets of data are used from March 1, 2003 to March 7, 2007. 22872 sets of data are used as training set, 8760 sets are defined as verification set, and 3600 sets are used as test set. The detailed characteristics of the two datasets are shown in Table 3, including total sample size, mean value, standard deviation, maximum and minimum.

B. EXPERIMENTAL SETTINGS

In order to verify the superior performance, the proposed model will be compared with the following models: rough autoencoder (RAE), DBN, interval probability distribution learning (IPDL), deep temporal dictionary learning (DTDL), TCN, LSTM, Bi-LSTM, iTCN, TCN-Attention (TCN-A), iTCN-A, CNN-LSTM-A, DenseNet-LSTM-A and DenseNet-TCN-A. All these models mentioned above will be run in the Python 3.7 environment using Pytorch 1.7 as back ends. The hardware is AMD 5800X CPU @3.80GHz and NVIDIA RTX3060 12GB GPU. The parameters of all these models were selected based on grid search optimization algorithm and previous experiences considering that these models required more parameter tuning skills. The parameters of each model are summarized as follow:

·RAE: The learning rate is 0.005, batch size is 512.

·DBN: The learning rate is 0.005, batch size is 512, the number of hidden layer is 10, the optimizer is Adam.

·DTDL: The dropout is 0.2, the size of the convolutional kernel is 3.

TABLE 3. The detailed characteristics of two public datasets.

Dataset		Amount	Statistic Data			
			Mean Value	Standard Deviation	Maximum	Minimum
Southern China	Total	55000	6614.08	2035.82	11404.36	1306.08
	Training Set	42848	6397.96	2025.22	11106.78	1306.08
	Validation Set	7152	6903.86	1767.98	10900.20	2349.97
	Testing Set	5000	8051.66	1839.50	11404.36	2860.66
ISO-NE	Total	35232	14940.37	2947.40	27622	8820
	Training Set	22872	14937.98	3003.98	26416	8820
	Validation Set	8760	14947.07	2957.78	27622	9018
	Testing Set	3600	14938.07	2529.21	21321	9525

TABLE 4. Load forecasting evaluation on the testing set.

Dataset		RAE	DBN	DTDL	IPDL	TCN	LSTM	Bi-LSTM	iTCN	TCN-A	iTCN-A	CNN-LSTM-A	DenseNet-LSTM-A	DenseNet-TCN-A	DenseNet-iTCN-A
Southern China	MAPE (%)	4.71	3.35	2.24	2.15	2.01	2.18	2.12	1.54	1.58	1.46	1.42	1.38	1.08	0.91
	MAE (MW)	582.06	464.26	354.56	312.54	278.02	318.11	170.55	231.43	233.86	201.31	221.43	133.23	153.52	121.28
	RMSE(MW)	986.62	814.96	579.54	511.97	414.46	500.62	248.16	308.11	333.22	258.03	328.44	203.1	215.14	176.01
	Cost time (s)	57.63	71.94	176.28	193.53	74.88	50.47	67.76	172.9	98.47	202.9	102.59	394.18	267.94	359.29
ISO-NE	MAPE (%)	4.07	2.57	1.90	1.72	1.35	1.41	1.45	1.24	1.19	1.13	1.09	1.03	0.95	0.87
	MAE (MW)	482.68	348.79	267.11	238.26	199.41	209.22	214.04	183.76	170.85	163.81	168.36	150.89	138.67	129.41
	RMSE(MW)	735.69	564.84	431.98	380.81	312.96	311.22	291.17	274.04	248.62	231.25	241.29	197.74	211.61	182.52
	Cost time (s)	48.54	58.17	123.98	142.25	59.25	43.29	52.32	117.51	62.28	138.96	70.41	111.41	139.42	170.85

·LSTM/Bi-LSTM: The number of hidden size is 24, the number of layers is 2, and the dropout is 0.2.

·CNN: The number of convolutional layers is 1, the size of the convolutional kernel is 3, and the stride is 1.

·TCN/iTCN: The dilation factor is set to [1, 2, 4, 8, 16], the kernel size is 2, the dropout is 0.2, the stride is 1, and the padding is 1.

·DenseNet: The number of dense block is 3, the size of the convolutional kernel is 3, and the stride is 1.

·Learning rate: The initial learning rate is set to 0.005, and the learning rate decay is adopted to automatically decrease the learning rate as the number of iterations increases. The learning rate is multiplied by 0.97 every 80 epochs.

C. RESULTS AND DISCUSSION

We performed the training and testing for all these models more than five times until the values of statistical metrics became stable. The mean values of experimental results of all these models on the two test sets are shown in Table 4. Figures 11 shows the MAPE errors of these aforementioned models in terms of a box plot on the testing sets of Southern China and ISO-NE datasets.

·Single model: Compared with the results of TCN, LSTM and Bi-LSTM, the MAPE of the iTCN decreased by 23.38%, 29.36%, 27.36% in Southern China dataset and 8.15%, 12.06%, 14.48% in ISO-NE dataset. This indicates that the iTCN has a stronger nonlinear fitting and prediction ability. It should pointed out that the first four machine learning models, i.e., RAE, DBN, DTDL and IPDL, also show inferior performance on these two datasets due to shallow structures of networks compared with other deep learning models.

·Self attention: The MAPE, MAE and RMSE of iTCN-A decreased by 5.19%, 13.01%, 16.25% in Southern China dataset and 8.8%, 10.8%, 15.6% in ISO-NE dataset. This fact is also true for the TCN and TCN-A. These comparison

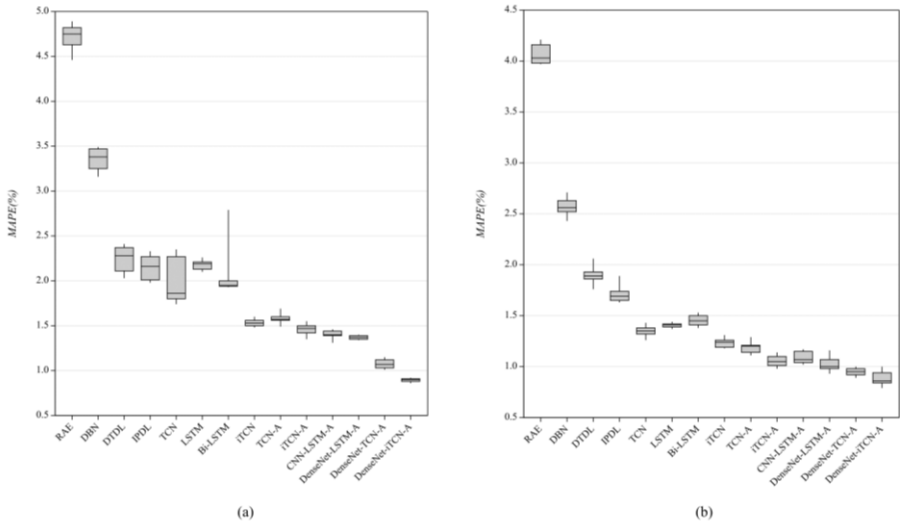

FIGURE 11. (a) The box diagrams of MAPE on Southern China dataset. (b) The box diagrams of MAPE on ISO-NE dataset.

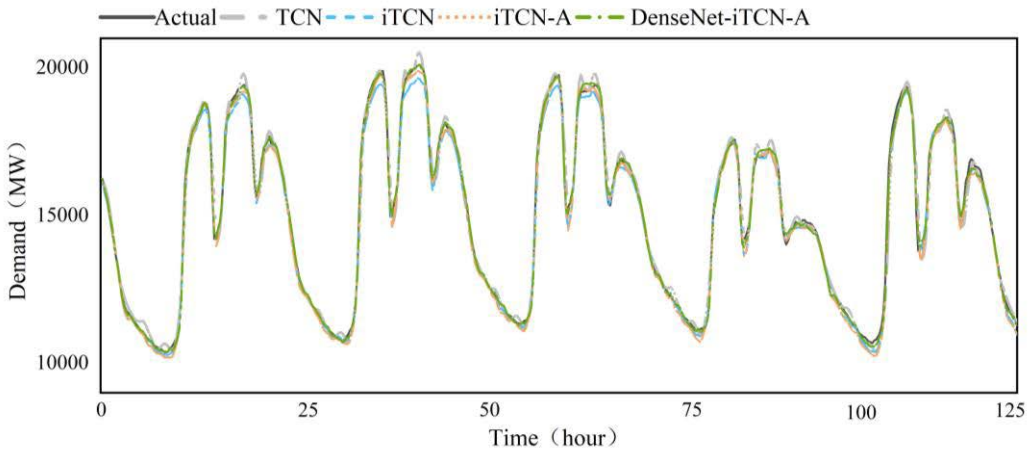

FIGURE 12. Load forecasting profiles of the TCN-based models of the dataset of Southern China.

results show that the SAM enhances the key features and further improves the learning ability of the prediction model.

·Hybrid model: Comparing CNN-LSTM-A with the proposed model, it can be found that all three evaluation metrics have been significantly reduced, especially for the values of MAE and RMSE. They are decreased by 35.92%, 45.23% and 46.41% in Southern China, 20.18%, 23.13% and 24.36% in ISO-NE. The reason is that the limitations of feature extraction technique lead to the loss of feature information. It means that the parallel pooling structure of the iTCN achieves a wider receptive field to capture the long-range historical data and get more detailed information of the features. Comparing DenseNet-LSTM-A and DenseNet-TCN-A with the proposed model on ISO-NE dataset, the MAPE are reduced by 15.53% and 8.42%, the MAE are reduced by 14.23% and 6.68, the RMSE are reduced by 7.70% and 13.75%. Thus, the proposed model is more significant and competitive than other hybrid models.

·Computational cost: It is obvious that hybrid models cost much more time than single models. The running time is positively proportional to the complexity of the model. It should stress that the hybrid models achieve higher accuracy compared to the single models. Thus, it is hard to balance the complexity of the model and prediction accuracy. However, it is acceptable in practical applications with the development of algorithms, graphics card and cloud computing. Table 5 shows the evaluation criteria of proposed model for 15-min, 30-min, 1-hour, and 2-hour ahead load forecasting, respectively. It is evident from Table 5 that the evaluation criteria are generally increased with the extension of the forecasting time horizon. The MAPE of 15-min obtained 16.50%, 34.35% and 44.87% improvement compared with 30-min, 1-hour, and 2-hour time steps, respectively.

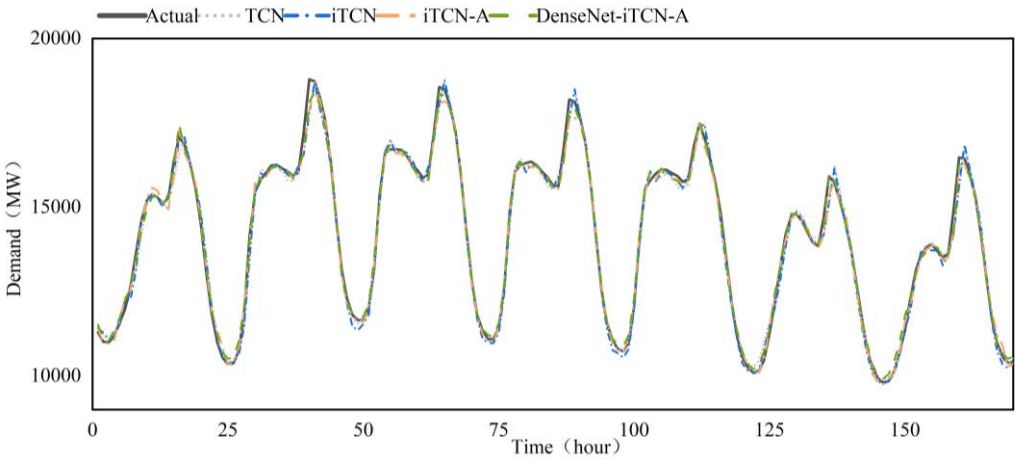

FIGURE 13. Load forecasting profiles of the TCN-based models of the dataset of ISO-NE.

TABLE 5. The criteria of proposed model for different time horizon.

	Time Step			
	15-min	30-min	1-hour	2-hour
MAPE(%)	0.86	1.03	1.31	1.56
MAE(MW)	117.15	146.68	206.69	232.50
RMSE(MW)	169.87	199.93	306.14	331.64

The forecasting results of several TCN-based models of the testing set, including the actual power demand curve, are shown in Figures12 and 13. It shows that all models can roughly fit the actual load in the stage of rising or falling. However, the proposed model is better able to fit and catch the trend of actual load compared to the deviations of other models at the peak or valley of the actual load. It is worth emphasizing that the ISO-NE dataset has lower complexity and non-stability compared to the dataset from a region in Southern China. Therefore, the errors of load forecasting for all models in the ISO-NE dataset are lower than those of corresponding models on Southern China dataset. In order to clearly present the comparison of all these models, we further show the load forecasting of 48 points (12 hours) on Southern China dataset in Figure 14 and the load forecasting of 24 points (24 hours) on ISO-NE dataset in Figure 15, respectively. It shows that all models can approximately fit the actual load in the stage of rising or falling. The peak or valley is not only the turning point of the curve but also the most difficult position for the model to predict. It should be pointed out that the peaks of the power load indicate the increase of the productivity and activities of residents and manufactories, which would be subjected to various disturbances. Similarly, the time horizons of valleys of the power load usually are very short between the falling and rising stages. In this case, some factors are unpredictable events, such as sudden load changes and others. All these factors will inevitably increase

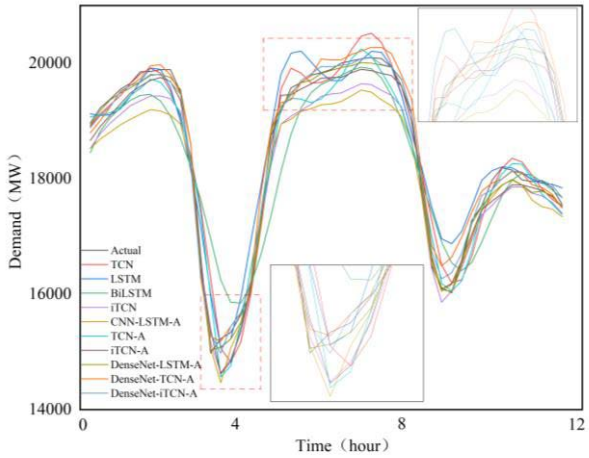

FIGURE 14. Load forecasting profiles of 12 hours on Southern China dataset.

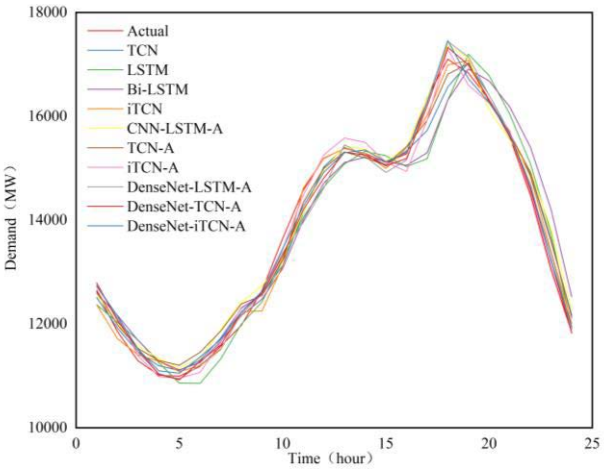

FIGURE 15. Load forecasting profiles of 24 hours on ISO-NE dataset.

the difficulty of load forecasting. Thus, the learning and prediction capability of the proposed model can be accurately

reflected due to the enhanced large receptive field and strong feature extraction and enhancement. It can be seen from the fitting curves that the proposed model in this study can better fit the changing trend of the actual power load. Therefore, we can conclude that our proposed method has high accuracy and robustness to meet the requirements of STLF.

IV. CONCLUSION

In this paper, we proposed a novel approach to short-term power load forecasting based on the DenseNet-iTCN and SAM. The information of features of the raw dataset was analyzed by correlation analysis method and the feature matrix is constructed as the input of the DenseNet to extract the in-depth features. A parallel pooling algorithm was introduced to the residual module of the TCN to effectively extract the temporal relationship of features without any signs of performance degradation or over-fitting. The key features separately obtained from the DenseNet and iTCN were enhanced via the self attention mechanism. The results of STLF with the optimal parameters were obtained through the full connection layer. Two public datasets from a region in Southern China and ISO-NE were performed to evaluate the proposed hybrid model. Experimental results showed that the proposed model had the best performance compared with other benchmarking models and demonstrated the strong generalization and robustness ability to STLF.

In future work, we will focus on specific areas of load forecasting, such as factories, communities, schools, hospitals, and other places. We will further optimize and improve the structure of TCN-based model to better adapt to time series.

REFERENCES

[1] M. Gilanifar, H. Wang, L. M. K. Sriram, E. E. Ozguven, and R. Arghandeh, "Multitask Bayesian spatiotemporal Gaussian processes for short-term load forecasting," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5132–5143, Jun. 2020.

[2] D. K. Ranaweera, G. G. Karady, and R. G. Farmer, "Economic impact analysis of load forecasting," *IEEE Trans. Power Syst.*, vol. 12, pp. 1388–1392, Aug. 1997.

[3] H. Shi, L. Wang, R. Scherer, M. Wozniak, P. Zhang, and W. Wei, "Short-term load forecasting based on adabelief optimized temporal convolutional network and gated recurrent unit hybrid neural network," *IEEE Access*, vol. 9, pp. 66965–66981, 2021.

[4] W. Lin, D. Wu, and B. Boulet, "Spatial-temporal residential short-term load forecasting via graph neural networks," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5373–5384, Nov. 2021.

[5] J. W. Taylor, "Short-term load forecasting with exponentially weighted methods," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 458–464, Feb. 2012.

[6] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: An evaluation based on European data," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 2213–2219, Nov. 2007.

[7] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Trans. Power Syst.*, vol. 16, no. 3, pp. 498–505, Aug. 2001.

[8] C.-M. Lee and C.-N. Ko, "Short-term load forecasting using lifting scheme and ARIMA models," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5902–5911, May 2011.

[9] B. Li, J. Zhang, Y. He, and Y. Wang, "Short-term load-forecasting method based on wavelet decomposition with second-order gray neural network model combined with ADF test," *IEEE Access*, vol. 5, pp. 16324–16331, 2017.

[10] C. Guan, P. B. Luh, L. D. Michel, and Z. Chi, "Hybrid Kalman filters for very short-term load forecasting and prediction interval estimation," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 3806–3817, Nov. 2013.

[11] S. Sharma, A. Majumdar, V. Elvira, and E. Chouzenoux, "Blind Kalman filtering for short-term load forecasting," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4916–4919, Nov. 2020.

[12] M. Khodayar and J. Wang, "Spatio-temporal graph deep neural network for short-term wind speed forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 670–681, Apr. 2019.

[13] S. Fan and L. Chen, "Short-term load forecasting based on an adaptive hybrid method," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 392–401, Feb. 2006.

[14] F. Liu, T. Dong, T. Hou, and Y. Liu, "A hybrid short-term load forecasting model based on improved fuzzy C-means clustering, random forest and deep neural networks," *IEEE Access*, vol. 9, pp. 59754–59765, 2021.

[15] C. Cecati, J. Kolbusz, P. Siano, and B. M. Wilamowski, "A novel RBF training algorithm for short-term electric load forecasting and comparative studies," *IEEE Trans. Ind. Electron.*, vol. 62, no. 10, pp. 6519–6529, Oct. 2015.

[16] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen, and C. Kang, "Combining probabilistic load forecasts," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3664–3674, Jul. 2019.

[17] J. W. Taylor and R. Buizza, "Neural network load forecasting with weather ensemble predictions," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 626–632, Aug. 2002.

[18] S. Afrasiabi, M. Afrasiabi, B. Parang, and M. Mohammadi, "Integration of accelerated deep neural network into power transformer differential protection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 865–876, Feb. 2020.

[19] M. Afrasiabi, M. Mohammadi, M. Rastegar, L. Stankovic, S. Afrasiabi, and M. Khazaei, "Deep-based conditional probability density function forecasting of residential loads," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3646–3657, Jul. 2020.

[20] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[21] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Sep. 2018.

[22] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[23] W. Wu, W. Liao, J. Miao, and G. Du, "Using gated recurrent unit network to forecast short-term load considering impact of electricity price," *Energy Proc.*, vol. 158, pp. 3369–3374, Feb. 2019.

[24] X. Kong, C. Li, F. Zheng, and C. Wang, "Improved deep belief network for short-term load forecasting considering demand-side management," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1531–1538, Mar. 2020.

[25] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He, "Short-term load forecasting with deep residual networks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3943–3952, Jul. 2019.

[26] X. Tang, H. Chen, W. Xiang, J. Yang, and M. Zou, "Short-term load forecasting using channel and temporal attention based temporal convolutional network," *Electric Power Syst. Res.*, vol. 205, Apr. 2022, Art. no. 107761.

[27] Y. Wang, J. Chen, X. Chen, X. Zeng, Y. Kong, S. Sun, Y. Guo, and Y. Liu, "Short-term load forecasting for industrial customers based on TCN-LightGBM," *IEEE Trans. Power Syst.*, vol. 36, no. 3, pp. 1984–1997, May 2021.

[28] P. Tang, P. Du, J. Xia, P. Zhang, and W. Zhang, "Channel attention-based temporal convolutional network for satellite image time series classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," Jun. 2018, *arXiv:1803.01271*.

[30] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1623–1631.

[31] H. Hao, Y. Wang, Y. Xia, J. Zhao, and F. Shen, "Temporal convolutional attention-based network for sequence modeling," Mar. 2020, *arXiv:2002.12530*.

[32] K. He, Z. Su, X. Tian, H. Yu, and M. Luo, "RUL prediction of wind turbine gearbox bearings based on self-calibration temporal convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[34] Q.-L. Hu, S.-L. Zhou, X.-G. Yu, G.-L. Xiao, X.-B. Luo, and R.-P. Cao, "Spin effects on the EM wave modes in magnetized plasmas," *Phys. Plasmas*, vol. 23, no. 11, Nov. 2016, Art. no. 112113.
[35] Y. Zhu, H. Luo, F. Zhao, and R. Chen, "Indoor/outdoor switching detection using multisensor DenseNet and LSTM," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1544–1556, Feb. 2021.
[36] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, "A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1407–1417, Jun. 2018.
[37] R. Cui and M. Liu, "Hippocampus analysis by combination of 3-D DenseNet and shapes for Alzheimer's disease diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2099–2107, Sep. 2019.
[38] I. Kreso, J. Krapac, and S. Segvic, "Efficient ladder-style DenseNets for semantic segmentation of large images," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4951–4961, Aug. 2021.
[39] K. Zhang, Y. Guo, X. Wang, J. Yuan, and Q. Ding, "Multiple feature reweight DenseNet for image classification," *IEEE Access*, vol. 7, pp. 9872–9880, 2019.
[40] Z. Zhu, G. Han, G. Jia, and L. Shu, "Modified DenseNet for automatic fabric defect detection with edge computing for minimizing latency," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9623–9636, Oct. 2020.
[41] L. Huang and C.-M. Pun, "Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1813–1825, Jul. 2020.
[42] C. Liu, L. Zhang, R. Yao, and C. Wu, "Dual attention-based temporal convolutional network for fault prognosis under time-varying operating conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.
[43] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, and J. Gall, "MS-TCN++: Multi-stage temporal convolutional network for action segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 4, 2020, doi: 10.1109/TPAMI.2020.3021756.
[44] D. Korkmaz, "SolarNet: A hybrid reliable model based on convolutional neural network and variational mode decomposition for hourly photovoltaic power forecasting," *Appl. Energy*, vol. 300, Oct. 2021, Art. no. 117410.
[45] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," Jul. 2016, *arXiv:1606.08415*.
[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[47] X. Zhang, G. Sun, X. Jia, L. Wu, A. Zhang, J. Ren, H. Fu, and Y. Yao, "Spectral–spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[48] J. Song and G. Xue, "Hourly heat load prediction model based on temporal convolutional neural network," *IEEE Access*, vol. 8, pp. 16726–16741, 2020.
[49] *Southern China Data Set*. Accessed: Dec. 6, 2020. [Online]. Available: https://github.com/keatoncu/Southern-China-Dataset
[50] *ISO-NE Data Set*. Accessed: Nov. 10, 2020. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/load-and-demand

MINGPING LIU received the B.Sc. and M.Sc. degrees from Jiangxi Normal University, Nanchang, China, and the Ph.D. degree from Beijing Normal University, Beijing, China. He is currently an Associate Professor with the School of Information Engineering, Nanchang University. His research interests include power system load forecasting, power system analysis, and deep learning.

HAO QIN received the B.Sc. degree from the Henan University of Technology, Henan, China, in 2020. He is currently pursuing the M.Sc. degree with the School of Information Engineering, Nanchang University, Nanchang, China. His research interests include power system load forecasting and deep learning.

RAN CAO received the B.Sc. degree from the Henan University of Technology, Henan, China, in 2020. She is currently pursuing the M.Sc. degree with the School of Information Engineering, Nanchang University, Nanchang, China. Her research interests include power system load forecasting and neural networks.

SUHUI DENG received the B.Sc. and M.Sc. degrees from Jiangxi Normal University, Nanchang, China, and the Ph.D. degree from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China. She is currently a Professor with the School of Information Engineering, Nanchang University. Her research interests include image processing, machine learning, and deep learning.