Power Load Forecasting Model Based on Improved Depthwise Separable Convolutional and Temporal Convolutional Network

Xue Zhu, Zhao Zhang, Hongyan Zhou, and Xue-Bo Chen

Abstract—**Short-term power load forecasts are designed to accurately predict electricity demand over the next few hours to days, to optimize the scheduling and operation of the power system. Forecasting accuracy is critical to the efficient dispatch and stable operation of the power system. To this end, this paper proposes a hybrid model for power load forecasting, which combines improved depthwise separable convolution and a temporal convolutional network. Specifically, this paper preprocesses the data, with missing values filled using the generative adversarial network. After data pre-processing, using the Pearson correlation coefficient to measure the degree of association of the data, and selecting appropriate exogenous variables. Long-term trends, seasonal fluctuations, and stochastic fluctuations in power load data are also extracted by seasonal decomposition algorithms. Secondly, the prediction model stacks multiple layers of depthwise separable convolutions, and the stacked depthwise separable convolutions are improved to incorporate an attention mechanism as well as residual connections, to perform feature extraction on power load data. Subsequently, time series features were processed in conjunction with a temporal convolutional network. Eventually, the prediction results are output through the fully connected layer. It is shown that the model outperforms other models in terms of load forecasting accuracy.**

Index Terms—**Power load forecasting, generative adversarial network, DSC, temporal convolutional network**

I. Introduction

POWER load forecasting is crucial in power system operation and management[1]. Power load forecasting relies on analyzing large-scale data, such as historical load records and meteorological conditions. Additionally, socio-economic factors, including population growth and economic development levels, must be considered. With a certain degree of accuracy, power load forecasting can predict future power usage for a given period[2]. Power load forecasting can be used in the day-to-day operations of utility companies (suppliers)[3]. For example, it can support scheduling power generation, power transmission, and real-time energy dispatch. Short Term Load Forecasting (STLF) is the basis for power dispatch and planning, helping power companies to develop rational power generation plans, optimize grid operations, and ensure reliability and economic supply[4-6]. With the growing demand for energy in modern society, the importance of power load forecasting is becoming more and more prominent[7]. Scientific forecasting of power loads has become a pressing issue for power companies and grid operators[8]. Effective management of electricity demand, along with reduced grid volatility and fewer power supply interruptions, helps to guarantee people's quality of life.

The accuracy of power load forecasting is critical to the efficient dispatch and stable operation of power systems. How to improve the accuracy of load forecasting has become the focus of research by scholars at home and abroad[9, 10]. New techniques and methods have been introduced at each stage of research development to improve the accuracy and reliability of forecasts. Earlier, statistical methods were widely used in power load forecasting. The foundation was laid by the AutoRegressive (AR) and Moving Average (MA) models proposed by George Box et al. On this basis, the Seasonal Autoregressive Integrated Moving Average Model (SARIMA) adds a seasonal component that can accommodate a variety of non-seasonal and seasonal patterns[11]. It can also be adapted to different characteristics of the time series, and this flexibility makes the SARIMA model suitable for a wide range of application scenarios[12, 13].

With the development of artificial intelligence and machine learning technologies, power load forecasting has entered a new phase[14]. Widely used models include Support Vector Machines (SVM), Artificial Neural Networks (ANN), and others. SVM models are particularly good at handling high-dimensional data and non-linear classification problems. Their main advantages include effectively capturing non-linear relationships and high-dimensional features, as well as maintaining high prediction accuracy with small sample sizes[15, 16]. ANN models are widely used in the field of artificial intelligence and machine learning.

Deep Learning (DL) is an important branch of artificial neural networks that emphasizes the use of multi-layer neural networks (deep neural networks, DNN) for feature extraction and representation learning[17]. Common deep learning models include Long Short-Term Memory (LSTM)

Manuscript received March 4, 2025; revised August 3, 2025. The research work was supported by the Fundamental Research Funds for the Liaoning Universities (LJ212410146025).

Xue Zhu is a graduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail: zhuxue.999@qq.com).

Zhao Zhang is an associate professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (corresponding author, e-mail: zhangzhao333@hotmail.com).

Hongyan Zhou is a doctoral student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail: zhou321yan@163.com).

Xue-Bo Chen is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (email: xuebochen@126.com).

Networks, introduced by Hochreiter et al., and their applications in time series forecasting. LSTM networks solve the problem of gradient vanishing and explosion in Recurrent Neural Networks (RNN) by introducing a gating mechanism that captures temporal dependencies in the data[18, 19]. The use of memory units to store important temporal information significantly improves the network's ability to handle long-term dependencies. In 2017, François Chollet introduced the concept of Depthwise Separable Convolution (DSC) and demonstrated its advantages in terms of computational efficiency and performance. DSC decomposes the convolution operation into depth convolution and point convolution. This decomposition reduces computational complexity, improves model efficiency, and maintains the performance of convolutional neural networks while reducing memory requirements. Also in 2017, Ashish Vaswani et al. introduced the self-attention mechanism of the Transformer model and demonstrated its effectiveness in processing long sequence data[20-24]. The Transformer is more efficient when dealing with long sequences because it does not depend on the previous moment's computation[25]. This independence allows Transformer to process long sequences faster and to parallelize them more easily.

To address the shortcomings of single models, Wolpert proposed the idea of multi-layer model integration. Zhou et al. provide a detailed overview of integration methods, describing how different models can be integrated to improve prediction performance. In recent years, much of the deep learning literature has begun to focus on combining different deep learning models to improve prediction. Hybrid predictive models overcome the limitations of a single model by combining the strengths of several different models, often resulting in better predictive performance in a variety of applications.

To improve the prediction performance of the model, an improved hybrid model of DSC and Temporal Convolutional Network (TCN) is proposed in this paper. The model captures patterns and features in a time series through efficient convolutional operations. DSC consists of two parts: depth convolution and point convolution. Depth convolution operates independently on each time series channel, enabling efficient extraction of features in each dimension (e.g., temperature, humidity, historical loads, etc. in power load data). Point convolution linearly combines these features through a 1×1 convolution kernel, integrating information from different channels to enhance the expressive and predictive performance of the model.

In this paper, multiple layers of DSCs are stacked, and the stacked DSCs are improved to include an attention mechanism as well as residual connections. TCN efficiently captures dependencies of long time series by causal convolution and dilated convolution. The causal convolution ensures that the output timestep depends only on the current and previous timesteps, preventing information leakage. Dilated convolution expands the receptive field to capture dependencies over longer time scales[26]. TCN also uses zero padding to ensure that the length of the output sequence after convolution is the same as the length of the input sequence, which is suitable for tasks where the length of the time series needs to be maintained.

In addition, to deal with the continuous vacancies in the power load data, this paper utilizes Generative Adversarial Networks (GAN) to efficiently fill in the missing data and make the dataset more complete by generating realistic data complements[27]. Selection of exogenous variables is a critical step in power load forecasting; this is because these variables significantly affect the performance of the predictive model[28-30]. In this paper, Pearson Correlation Coefficient (PCC) is used as a valid statistical method for detecting and quantifying relationships between variables. In order to better understand the stochastic and cyclical characteristics of power loads[31], in this paper, a seasonal decomposition algorithm is used to decompose the power load data into trend, seasonal, and residual components. This decomposition allows for deeper learning and modelling of the different characteristics of power loads, thus improving the accuracy of the forecasting model.

The main innovations and contributions of this paper are as follows:

1) In this paper, an innovative power load forecasting model is proposed, which combines improved DSC and TCN. The improved DSC excels in feature

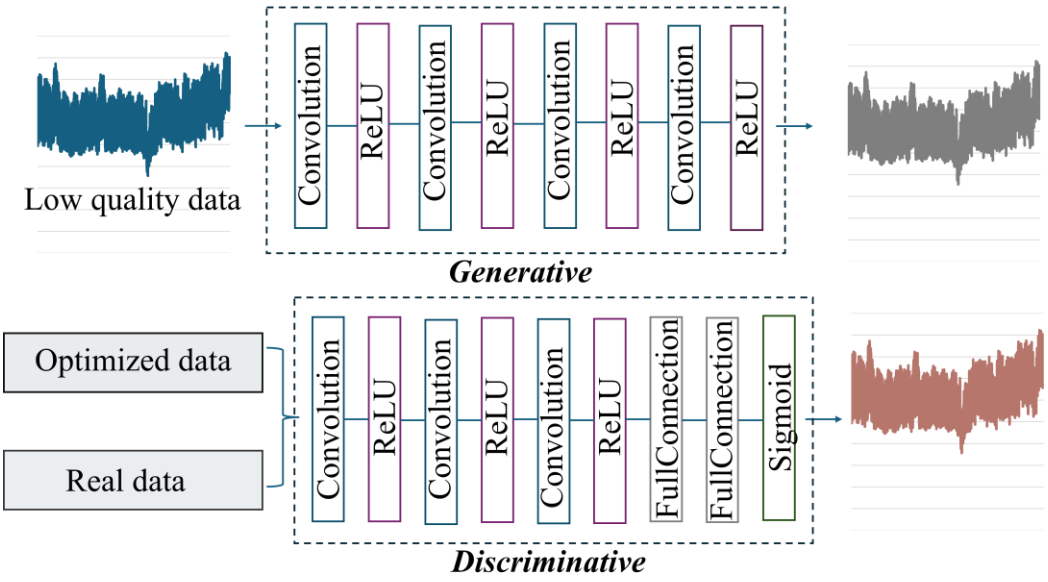

Fig 1. GAN structure

extraction and computational optimization, while the TCN excels in modelling long time dependencies. This combination not only improves the model's prediction performance on time series data but also achieves a good balance between computational efficiency and processing power. Such models possess efficient feature extraction and processing capabilities, enhanced capacity to model short- and long-term dependencies, and strike a balance between computational efficiency and model complexity.

2) To fill the gaps in data, in this paper, GAN is used to deal with the problem of missing power load data. It takes better account of the complex relationships and dependencies between data than traditional methods.
3) In this paper, exogenous variables are selected using the Pearson correlation coefficient and then combined with seasonal decomposition results as inputs to the time series forecasting model. This preprocessing step provides the model with highly relevant and denoised input data.

II. Date Processing

A. Data Preprocessing

Missing data may have a significant impact on power load forecasting. Missing data can lead to incomplete data being used by the model during training, thereby affecting the learning of data patterns and features by the model. Inaccurate forecasts may result from models that do not fully capture trends and seasonal variations in the data. This may also make the model training process more difficult. Filling in missing data helps to maintain data continuity and integrity; this is essential for the model to learn patterns and trends in the data.

In this paper, GAN is used to fill in continuous data. As shown in Fig. 1, GAN fills in the missing data by generating values that are close to real data, thereby recovering the incomplete parts and improving the accuracy of the predictive model.

In the training process, the generator and discriminator are applied together. The generator uses the Mean Square Error (MSE) loss function to generate data that are as realistic as possible, making it difficult for the discriminator to distinguish between generated and real data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i is the actual value of the i-th sample, $\hat{y}_i$ is the predicted value of the i-th sample, n is the total number of samples.

The discriminator optimizes its parameters using the binary classification crossentropy loss function V. The goal is to distinguish as accurately as possible between real and generated data.

$$V(D,G)=\frac{1}{\mathrm{m}}\sum_{i=1}^{m}[\log D(x_i)+\log(1-D(G(z_i)))] \tag{2}$$

where m denotes a total of m samples. x_i denotes arbitrary real data, z_i denotes arbitrary random data with the same structure as the real data. $G(\mathrm{z}_i)$ denotes the fake data generated in the generator based on z_i, $D(\mathrm{x}_i)$ denotes the result judged by the discriminator on the real data x_i. $D(G(z_i))$ denotes the result judged by the discriminator on the false data $G(z_i)$. Where $D(\mathrm{x}_i)$ and $D(G(z_i))$ are the probability that the sample is true. The loss V reaches its maximum value when the judgment of the discriminator is at its best.

The trained generator can be used to fill in data with missing values in the test set. The filler data generated by the generator replaces missing values, thus recovering the complete data.

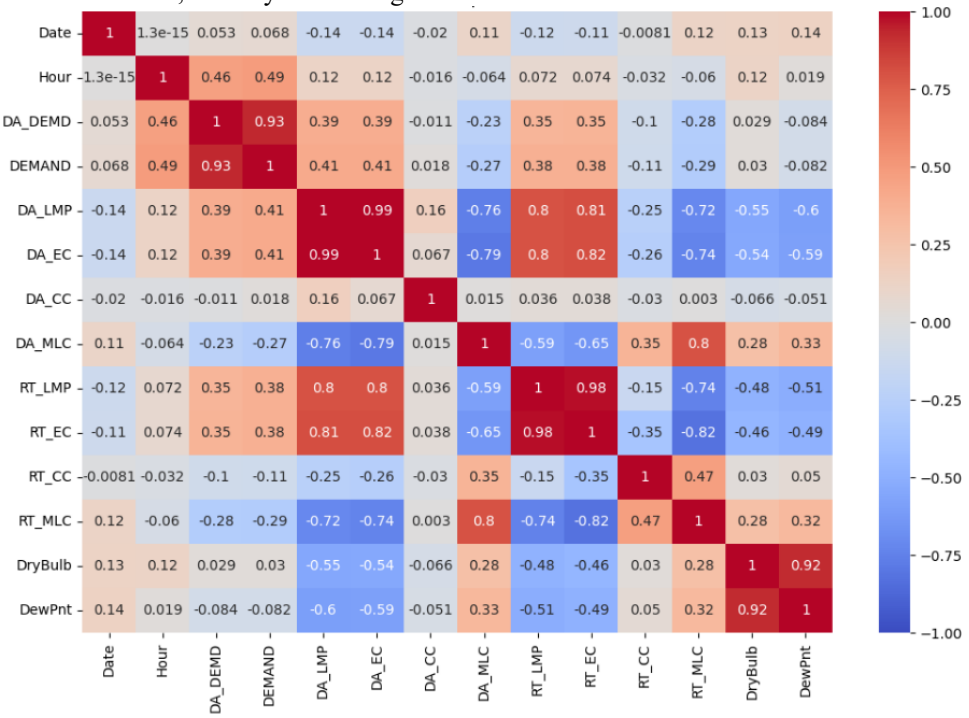

Fig 2. Heat map of Pearson correlation coefficient between power load and characteristic variables

B. Data Characterisation

Choosing the right features can help models capture patterns and trends in the data more accurately and reduce noise interference, thereby improving the accuracy of forecasts[32]. Feature selection reduces the number of model input variables and helps identify which factors significantly impact load, thereby facilitating a better understanding of load patterns.

Complex non-linear relationships may exist between exogenous variable characteristics and target variables in power load data. In this paper, we choose to use the Pearson Correlation Coefficient (PCC) to effectively capture these linear and non-linear dependencies, which helps identify the most relevant exogenous variables.

$$r = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} \cdot \sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{3}$$

where r is the Pearson correlation coefficient, Range $[-1,1]$. a_i , b_i are the observed values of variables A , B . $\bar{a}$, $\bar{b}$ are the mean values of variables A , B .

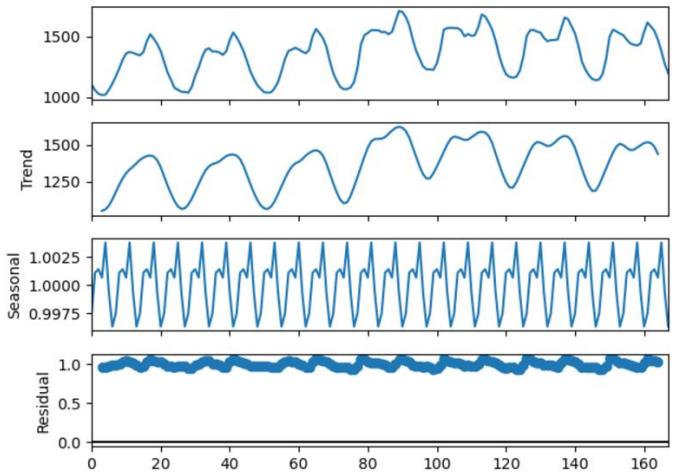

Fig 3. Seasonal decomposition algorithm decomposition diagram

The heat map of the Pearson correlation coefficient between load and characteristic variables is shown in Fig. 2.

The PCC is used to measure the relationship between power load and external variables such as temperature and weather, and to select external variables. While choosing the right exogenous variables, the decomposition of time series data y_t into trend $T(t)$, seasonal $S(t)$ and stochastic components $R(t)$ using seasonal decomposition algorithm. As shown in Fig. 3. This can help identify long-term trends and cyclical fluctuations, thus better understanding of the changing law of power load.

$$y(t) = T(t) \times S(t) \times R(t) \tag{4}$$

$$S(t) = \frac{y(t)}{T(t)} \tag{5}$$

$$R(t) = \frac{y(t)}{T(t) \times S(t)} \tag{6}$$

The overall process is shown in Fig. 4, where the data is processed and fed into the prediction model.

III. Power Load Forecasting Model

The improved hybrid model of DSC and TCN is described in detail in this section.

The data is processed and fed into a predictive model. In this paper, we use a power load forecasting model consisting of multiple layers of improved depthwise separable convolutional layers[33], a temporal convolutional network layer, and a fully connected layer. The fully connected layer outputs the predicted power load for the next 24 hours. The improved DSC-TCN model is shown in Fig. 5.

Specifically, the improved DSC, first, the input feature Y is subjected to a DSC operation, as shown in Fig. 6. The DSC contains depth convolution and point convolution. Depth convolution performs the convolution operation independently for each input channel, unlike conventional

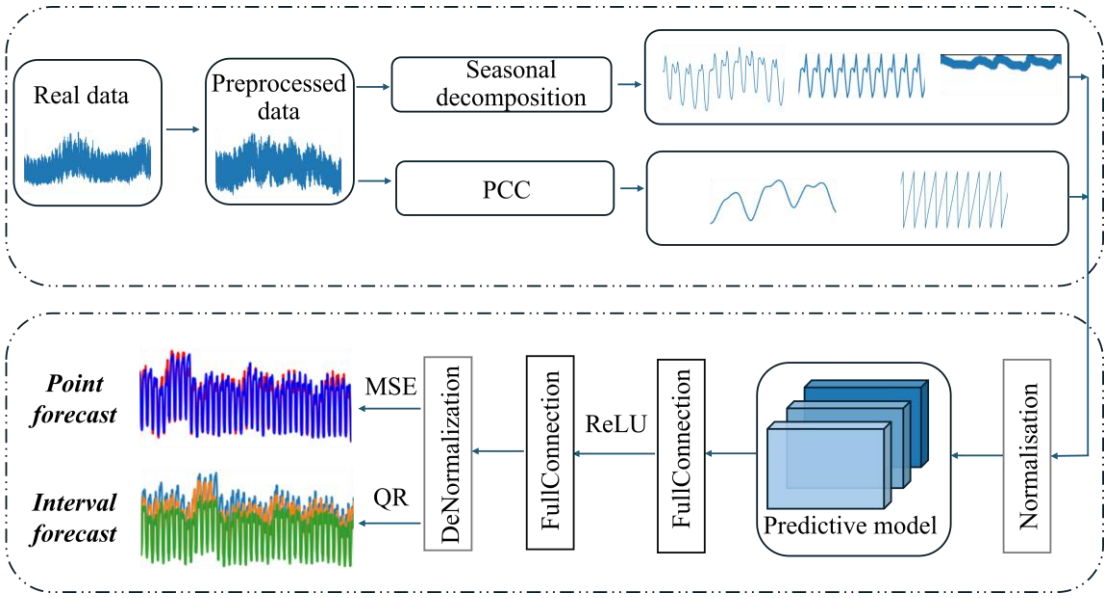

Fig 4. Overall flow chart

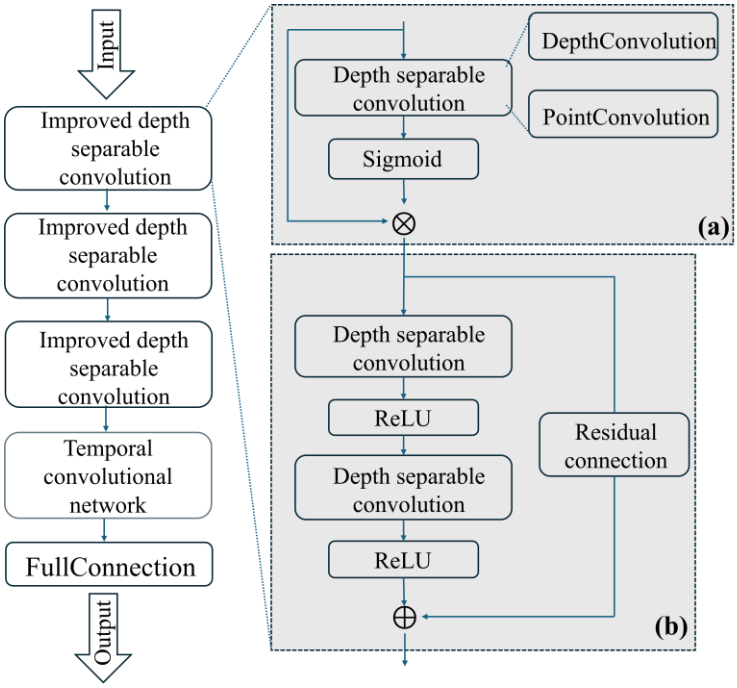

Fig 5. Structural diagram of the improved DSC-TCN model

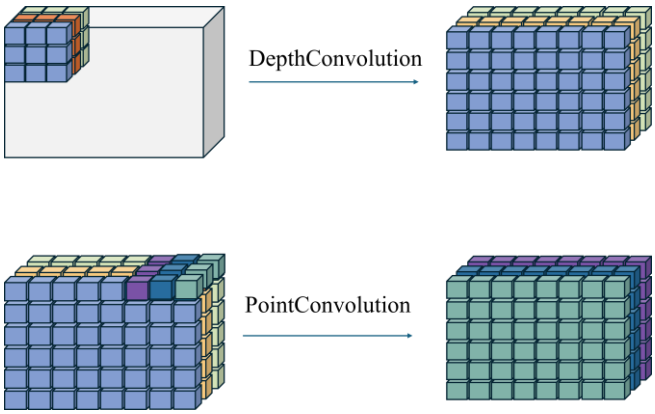

Fig 6. Depth separable convolutional structure diagram

convolution which convolves all input channels. Then, point convolution uses a 1×1 convolution kernel to convolve the output of the depth convolution, to fuse channel information and change the number of channels.

The DSC output features α go through the Sigmoid activation function, generating attention weights λ, these weights indicate the importance of each position.

$$\lambda = Sigmoid(\alpha) = \frac{1}{1+e^{-\alpha}} \tag{7}$$

Using these attentional weights, the original input features are adjusted to highlight the important parts and assign weights to input time steps and input features. The output of the feature after adding the attention mechanism is β_1.

$$\beta_1 = \lambda \cdot Y \tag{8}$$

More complex and abstract features can be extracted incrementally using multi-layer DSCs, enabling the model to capture more detailed patterns and information. By processing layer by layer, the model can continuously optimize the feature representation and enhance both its nonlinear capabilities and overall expressiveness. The weighted features are then processed through two layers of DSC to obtain the output feature β_3 [33].

$$\begin{aligned} \beta_2 &= \mathrm{Re}\,LU(DSC_2(\beta_1)) \\ \beta_3 &= \mathrm{Re}\,LU(DSC_3(\beta_2)) \end{aligned} \tag{9}$$

At the same time, the inclusion of residual connections in the model makes the training of deep networks more stable and efficient, enabling the network to optimize and converge more easily. Attention mechanisms are included in the input features of residual connections. Sum the input with the processed output features, for the ultimate in improved DSC output characteristics β_{final}.

$$\beta_{final} = \beta_3 + Residual \tag{10}$$

In the residual connection, when the shape of input feature β_1 matches the shape of output feature β_3, they can be directly summed; When the shape of β_1 does not match the shape of β_3, a convolution operation is needed to adjust the shape of the input features, align it with the output features.

$$Residual = \begin{cases} \beta_1, \beta_1 = \beta_3 \\ Stack_N(W_{conv} * \beta_1 + k_{conv}), \beta_1 \neq \beta_3 \end{cases} \tag{11}$$

where W_{conv} denotes the convolution kernel, $*$ denotes the convolution operation and k_{conv} is the bias.

In this paper, three improved depth-separable convolutional layers are used, and the model performance can be improved.

Then, TCN was applied for effective processing and feature extraction of time series data. TCN contains causal convolution and dilated convolution. TCN uses causal convolution to ensure that only current and past time-step data are used, thus avoiding information leakage. The causal convolution formula is:

$$C(t) = \sum_{j=0}^{J-1} \varpi(j) \cdot u(t-j) \tag{12}$$

where $u(t-j)$ denotes the input value at time step $t-j$, $\varpi(j)$ denotes the weight of the convolution kernel at

position j, J denotes the length of the convolution kernel. By dilated convolution, TCN extends the sensory field, capable of capturing long distance dependencies, while maintaining computational efficiency. The dilated convolution formula is:

$$C(t)=\sum_{j=0}^{J-1}\varpi(j)\cdot u(t-d\cdot j) \tag{13}$$

where d is the expansion rate, which controls the spacing of the convolution kernels. In this paper the first layer expansion is 1, the second expansion is 2, and the third layer expansion is 4. $t-d\cdot j$ is the input index of the expansion convolution.

Meanwhile, the use of residual connections in TCN further enhances the stability of the network. Reduced the problem of disappearing gradients, which makes the model training process smoother and more efficient.

IV. Model Performance Analysis

A. Experimental Environment

The configuration of the experimental environment is shown in Table I.

TABLE I
EXPERIMENTAL ENVIRONMENT CONFIGURATION TABLE

Environment	Parametric	Configure
Hardware Environment	CPU	AMD Ryzen 9 7950X 16-Core Processor 4.50 GHz
	RAM	64GB
	GPU	GeForce RTX 4090 D 24G
Software Environment	operating system	Windows 11
	development framework	PyTorch 2.5.1
	development language	Python 3.9.20

B. Experimental Data

The data used in this paper are derived from power load demand data from 1 January 2013 to 31 December 2014 for the New England region of the United States, and from 1 January to 31 December 2018 for the Nordic region. The sampling interval for each historical load sample is 1 hour. The dataset includes multiple cities and regions, such as Connecticut (CT), Maine (ME), and New Hampshire (NH), and provides rich information on power loads and their associated features. The input features of this paper include two PCC-selected exogenous variables, three post-decomposition variables for power load data, and raw power load data. This paper uses the data from reference [33] for comparison, and the results demonstrate the accuracy of the proposed model.

C. Results

The improved DSC-TCN combined model demonstrates its significant prediction accuracy and computational efficiency advantages in power load forecasting. Compared to other models, the improved model has advantages in terms of long-range dependency capture, model stability and feature extraction capability. In this paper, we compare the point prediction and interval prediction results of the improved DSC-TCN combination model with the improved DSC model. The input in this paper are all 6 channels of input data.

Point Forecast

Power load point forecasts are forecasts of power use over the next 24 hours; this refers to the forecast of a single value of power load at a future point in time. This forecast provides a power load value that indicates the expected demand, supporting planning and scheduling to ensure a stable and reliable power supply. In this paper, MSE is used as a loss function to obtain predictive data. The MSE formula is shown in Eq. (1).

Three evaluation criteria Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R Squared (R2) were selected to analyze the results of the point prediction[34]. As shown in Table II, From the three evaluation criteria in the table, it can be seen that the point prediction performance of the improved DSC-TCN combined model is better than that of the improved DSC model.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|p_i-\hat{p}_i| \tag{14}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-\hat{p}_i)^2} \tag{15}$$

$$R2=1-\frac{\sum_{i=1}^{n}(p_i-\hat{p}_i)^2}{\sum_{i=1}^{n}(p_i-\bar{p})^2} \tag{16}$$

where n is the number of samples, p_i is the actual value of the i-th sample, $\hat{p}_i$ is the predicted value for the i-th sample, $\bar{p}$ is the mean of the actual values.

when $R2 = 1$ indicates that the model fits the data perfectly, all variation is explained by the model.

As can be seen in Table II, the improved combined DSC-TCN model shows different levels of improvement for different evaluation metrics of point prediction. Fig. 7 illustrates the predicted results of the model point predictions in a graph, only one section of the data was selected.

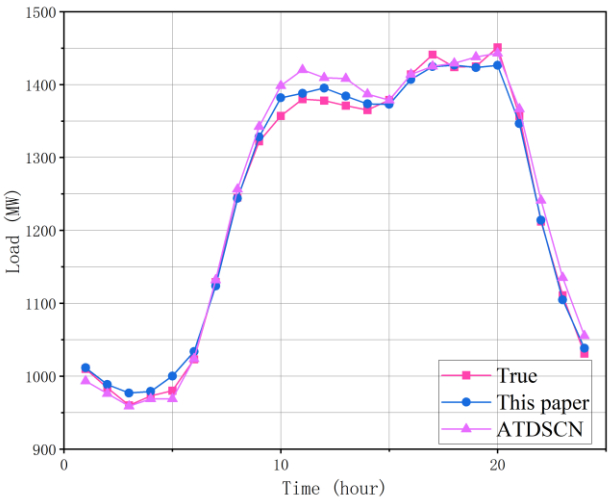

Fig 7. Point prediction results of the improved DSC-TCN model

Interval Forecast

Interval prediction provides a prediction interval, not a single value, indicates that future load values may fall within this range. Power load interval forecasts can provide upper and lower bounds on forecast values, giving the range of

TABLE II
INDICATORS FOR ASSESSING MODEL POINT PROJECTIONS

Region		FNN	GRU	Transformer	CNN-LSTM	TCN	CNNs-Transformer	DR-DNN	AM-LSTM	ATDSCN[33]	**This paper**
CT	MAE	130.1	136.5	141.4	143.6	134.3	135	133.9	141.9	121.8	**107.4**
	RMSE	197	204.5	207.1	217.9	201.6	199.5	201.9	215.5	188.1	**150.9**
	R2	0.912	0.908	0.901	0.896	0.907	0.91	0.908	0.898	0.921	**0.951**
ME	MAE	33.66	36.37	38.99	40.44	35.19	36.65	35.89	40.11	32.78	**25.81**
	RMSE	49.7	52.28	54.98	56.51	51.67	52.69	52.54	56.38	48.93	**34.19**
	R2	0.929	0.924	0.915	0.912	0.921	0.921	0.922	0.914	0.932	**0.972**
VT	MAE	20.45	21.35	22.62	22.92	21.45	21.76	21.09	22.44	19.72	**15.45**
	RMSE	29.29	30.37	31.55	32.15	30.43	30.46	30	31.71	28.44	**21.3**
	R2	0.909	0.906	0.898	0.896	0.9	0.903	0.906	0.897	0.915	**0.962**
RI	MAE	33.8	33.84	37.49	37.57	34.57	35.08	34.84	37.26	31.87	**29.96**
	RMSE	51.01	52.93	54.45	57.43	51.81	52.76	51.55	57.52	48.99	**42.13**
	R2	0.912	0.91	0.901	0.893	0.908	0.909	0.911	0.895	0.92	**0.942**
NH	MAE	43.61	42.85	49.22	48.69	44.83	45.89	43.86	46.42	41.4	**34.64**
	RMSE	68.31	67.01	74.58	75.77	70.54	70.37	68.47	72.56	66.76	**48.73**
	R2	0.92	0.926	0.908	0.905	0.913	0.917	0.921	0.912	0.925	**0.963**
EE	MAE	27.85	30.29	33.38	34.24	27.99	31.49	27.86	35.56	27.86	**31.04**
	RMSE	42.92	43.6	48.63	48.92	41.14	46.52	41.77	51.75	41.77	**40.68**
	R2	0.944	0.941	0.93	0.928	0.948	0.935	0.946	0.918	0.946	**0.955**
DK	MAE	68.38	75.29	76.75	86.33	69.04	72.88	69.88	85.11	69.88	**67.37**
	RMSE	107.1	114.6	125.4	132.5	105.7	112	108.6	131.7	108.6	**85.41**
	R2	0.941	0.933	0.922	0.904	0.941	0.937	0.94	0.91	0.94	**0.965**
LV	MAE	22.18	26.45	24.81	27.9	22.52	24.63	21.69	28.39	21.69	**23.63**
	RMSE	37.06	40.57	41.64	43.31	36.91	40.58	37.08	43.9	37.08	**34.13**
	R2	0.953	0.944	0.945	0.934	0.953	0.945	0.954	0.934	0.954	**0.963**

TABLE III
INDICATORS FOR ASSESSING MODEL INTERVAL PROJECTIONS

Region		FNN	GRU	Transformer	CNN-LSTM	TCN	CNNs-Transformer	DR-DNN	AM-LSTM	ATDSCN[33]	**This paper**
CT	CWC	0.221	0.223	0.241	0.232	0.226	0.236	0.235	0.222	0.214	**0.185**
	RWS	0.12	0.119	0.131	0.122	0.126	0.126	0.128	0.116	0.119	**0.090**
ME	CWC	0.213	0.219	0.215	0.234	0.219	0.218	0.217	0.225	0.205	**0.201**
	RWS	0.092	0.09	0.088	0.094	0.096	0.091	0.092	0.09	0.087	**0.064**
VT	CWC	0.244	0.254	0.265	0.265	0.263	0.263	0.247	0.279	0.242	**0.213**
	RWS	0.094	0.094	0.099	0.098	0.102	0.1	0.094	0.104	0.093	**0.076**
RI	CWC	0.219	0.213	0.245	0.231	0.225	0.228	0.23	0.218	0.208	**0.207**
	RWS	0.117	0.111	0.126	0.118	0.119	0.119	0.122	0.111	0.111	**0.103**
NH	CWC	0.224	0.221	0.246	0.237	0.235	0.238	0.234	0.222	0.217	**0.201**
	RWS	0.108	0.101	0.114	0.107	0.115	0.11	0.109	0.102	0.103	**0.087**
EE	CWC	0.248	0.286	0.289	0.319	0.274	0.299	0.254	0.311	0.246	**0.245**
	RWS	0.081	0.089	0.088	0.096	0.092	0.093	0.082	0.095	0.081	**0.085**
DK	CWC	0.261	0.292	0.278	0.314	0.267	0.266	0.268	0.275	0.246	**0.268**
	RWS	0.095	0.1	0.096	0.107	0.097	0.096	0.099	0.097	0.088	**0.088**
LV	CWC	0.212	0.238	0.227	0.285	0.231	0.24	0.212	0.272	0.203	**0.205**
	RWS	0.075	0.074	0.074	0.091	0.078	0.078	0.073	0.088	0.068	**0.064**

uncertainty of the load. In this paper, Quantile Regression (QR) is used as the loss function, selecting the desired quantile for defining the lower and upper bounds of the interval respectively, to define a prediction interval. The loss function formula is shown in Eq. (17).

$$L\tau(q,\hat{q}) = \tau \cdot \max(q-\hat{q},0) + (1-\tau)\cdot \max(\hat{q}-q,0) \quad (17)$$

where q is the true value, $\hat{q}$ is the predicted value of the model. τ is the quantile parameter, its value ranges between $(0,1)$ and in this paper it takes the value of 0.05.

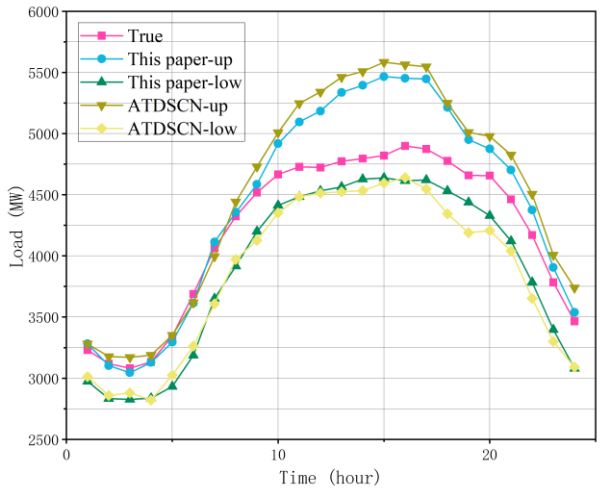

Fig 8. Interval prediction results of the improved DSC-TCN model

In this paper, the evaluation criteria of Coverage Width Criterion (CWC) and Relative Width Score (RWS) were selected to analyze the interval prediction results. The CWC combines forecast interval width and coverage, is an indicator that balances breadth and accuracy. A lower CWC indicates that the prediction interval is as narrow as possible while maintaining high coverage and the model predicts better. RWS assesses the tightness and accuracy of the model's prediction intervals. A smaller value of RWS indicates a narrower prediction interval with guaranteed coverage of the prediction interval, and more accurate prediction intervals.

$$CWC = \frac{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(U_i - L_i)}{R} \cdot (1 + \gamma \cdot e^{-\beta \cdot (PICP - \mu)}) \tag{18}$$

$$\gamma = \begin{cases} 0, PICP \geq \mu \\ 1, PICP \leq \mu \end{cases} \tag{19}$$

$$PICP = \frac{\sum_{i=1}^{n} \varphi(L_i \leq q_{true,i} \leq U_i)}{\mathrm{n}} \tag{20}$$

where U_i , L_i are the upper and lower bounds of the prediction interval; R is the target value range. n is the sample size; γ is the penalty factor; β is the penalty factor, and the value of β in the article is 2. β is used to control the intensity of the penalty when $PICP < \mu$, μ is 0.9 in the article. $PICP$ is the prediction interval coverage probability. where $q_{true,i}$ is the true value. $\varphi(x)$ is the indicator function, takes a value of 1 when the condition is true, and 0 otherwise.

$$\delta_{\mathrm{i}} = 2 \cdot \frac{U_i - L_i}{U_i + L_{\mathrm{i}}} \tag{21}$$

$$L\_RWS_{\mathrm{i}} = \begin{cases} \delta_i, L_i \leq q_{true,i} \leq U_i \\ \delta_i + \dfrac{L_i - q_{true,i}}{q_{true,i}}, q_{true,i} < L_i \\ \delta_i + \dfrac{q_{true,i} - U_i}{q_{true,i}}, q_{true,i} > U_i \end{cases} \tag{22}$$

$$RWS = \frac{\sum_{i=1}^{n} L\_RWS_i}{n} \tag{23}$$

where δ_i is the relative width of the prediction interval, measures the size of the interval width relative to its position. L_RWS_{i} is the relative score for each sample, including interval widths and penalties for exceeding the interval.

Table III compares the model interval prediction assessment metrics CWC and RWS. From the evaluation criteria in the table, it can be seen that the improved DSC-TCN combination model interval prediction performance is better than the improved DSC model. Fig. 8 illustrates a plot of the prediction results for the model interval prediction, where only one segment of data was selected.

In summary, the model in this paper shows good prediction results in both point prediction and interval prediction.

V. Conclusion

Power load forecasting is crucial for the stability and sustainability of the electricity supply. Traditional methods often struggle with missing or inaccurate historical data and fail to adequately consider external factors like climate change and seasonal variations. This paper addresses these issues by using GAN to fill in missing data, PCC to select exogenous variables, and seasonal decomposition to reduce noise and capture seasonal features. An improved DSC-TCN model is proposed, combining local feature extraction and long-term time-dependent modeling for more accurate forecasts. Although the model performs well, there are some limitations including the fact that the GAN requires a large training dataset and that the PCC may be inaccurate due to outliers, which are areas for future improvement.

References

[1] T. Li, Y. Wang, and N. Zhang, "Combining Probability Density Forecasts for Power Electrical Loads," *IEEE Transactions on Smart Grid,* vol. 11, no. 2, pp. 1679-1690, 2020.

[2] K. Li, W. Huang, G. Hu, and J. Li, "Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network," *Energy and Buildings,* vol. 279, 2023.

[3] I. K. Nti, M. Teimeh, O. Nyarko-Boateng, and A. F. Adekoya, "Electricity load forecasting: a systematic review," *Journal of Electrical Systems and Information Technology,* vol. 7, no. 1, 2020.

[4] A. Wan, Q. Chang, K. Al-Bukhaiti, and J. He, "Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism," *Energy,* vol. 282, 2023.

[5] J. Zhang, Z. Tan, and S. Yang, "Day-ahead electricity price forecasting by a new hybrid method," *Computers & Industrial Engineering,* vol. 63, no. 3, pp. 695-701, 2012.

[6] Y. Chen, and H. Tan, "Short-term prediction of electric demand in building sector via hybrid support vector regression," *Applied Energy,* vol. 204, pp. 1363-1374, 2017.

[7] Z. Xia, R. Zhang, H. Ma, and T. K. Saha, "Day-Ahead Electricity Consumption Prediction of Individual Household–Capturing Peak Consumption Pattern," *IEEE Transactions on Smart Grid,* vol. 15, no. 3, pp. 2971-2984, 2024.

[8] R. Sarwar, H. Cho, S. J. Cox, P. J. Mago, and R. Luck, "Field validation study of a time and temperature indexed autoregressive with exogenous (ARX) model for building thermal load prediction," *Energy,* vol. 119, pp. 483-496, 2017.

[9] Z. Guo, K. Zhou, X. Zhang, and S. Yang, "A deep learning model for short-term power load and probability density forecasting," *Energy,* vol. 160, pp. 1186-1200, 2018.

[10] L. Zhang, J. Shi, L. Wang, and C. Xu, "Electricity, Heat, and Gas Load Forecasting Based on Deep Multitask Learning in Industrial-Park Integrated Energy System," *Entropy,* vol. 22, no. 12, 2020.

[11] J. Naik, P. Satapathy, and P. K. Dash, "Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression," *Applied Soft Computing,* vol. 70, pp. 1167-1188, 2018.
[12] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 3125-3148, 2019.
[13] S. Chaturvedi, E. Rajasekar, S. Natarajan, and N. McCullen, "A comparative assessment of SARIMA, LSTM RNN and Fb Prophet models to forecast total and peak monthly energy demand for India," *Energy Policy,* vol. 168, 2022.
[14] H. Hua, M. Liu, Y. Li, S. Deng, and Q. Wang, "An ensemble framework for short-term load forecasting based on parallel CNN and GRU with improved ResNet," *Electric Power Systems Research,* vol. 216, 2023.
[15] M. Barman, and N. B. Dev Choudhury, "A similarity based hybrid GWO-SVM method of power system load forecasting for regional special event days in anomalous load situations in Assam, India," *Sustainable Cities and Society,* vol. 61, 2020.
[16] C. Tarmanini, N. Sarma, C. Gezegin, and O. Ozgonenel, "Short term load forecasting based on ARIMA and ANN approaches," *Energy Reports,* vol. 9, pp. 550-557, 2023.
[17] A. O. Aseeri, "Effective RNN-Based Forecasting Methodology Design for Improving Short-Term Power Load Forecasts: Application to Large-Scale Power-Grid Time Series," *Journal of Computational Science,* vol. 68, 2023.
[18] C. Peng, Y. Tao, Z. Chen, Y. Zhang, and X. Sun, "Multi-source transfer learning guided ensemble LSTM for building multi-load forecasting," *Expert Systems with Applications,* vol. 202, 2022.
[19] H. Zang, R. Xu, L. Cheng, T. Ding, L. Liu, Z. Wei *et al.*, "Residential load forecasting based on LSTM fusing self-attention mechanism with pooling," *Energy,* vol. 229, 2021.
[20] K. Denecke, R. May, and O. Rivera-Romero, "Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks," *Journal of Medical Systems,* vol. 48, no. 1, 2024.
[21] A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A Historical Survey of Advances in Transformer Architectures," *Applied Sciences,* vol. 14, no. 10, 2024.
[22] A. Rahali, and M. A. Akhloufi, "End-to-End Transformer-Based Models in Textual-Based NLP," *Ai,* vol. 4, no. 1, pp. 54-110, 2023.
[23] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," *Information,* vol. 14, no. 4, 2023.
[24] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 35, no. 12, pp. 11106-11115, 2021.
[25] Q. Zhang, J. Zhang, Y. Xu, and D. Tao, "Vision Transformer With Quadrangle Attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 46, no. 5, pp. 3608-3624, 2024.
[26] L. Yin, and J. Xie, "Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems," *Applied Energy,* vol. 283, 2021.
[27] X. Kang, L. Liu, and H. Ma, "ESR-GAN: Environmental Signal Reconstruction Learning With Generative Adversarial Network," *IEEE Internet of Things Journal,* vol. 8, no. 1, pp. 636-646, 2021.
[28] N. Kim, H. Park, J. Lee, and J. K. Choi, "Short-Term Electrical Load Forecasting With Multidimensional Feature Extraction," *IEEE Transactions on Smart Grid,* vol. 13, no. 4, pp. 2999-3013, 2022.
[29] N. Bashiri Behmiri, C. Fezzi, and F. Ravazzolo, "Incorporating air temperature into mid-term electricity load forecasting models using time-series regressions and neural networks," *Energy,* vol. 278, 2023.
[30] J. Wang, L. Han, X. Zhang, Y. Wang, and S. Zhang, "Electrical load forecasting based on variable T-distribution and dual attention mechanism," *Energy,* vol. 283, 2023.
[31] Z. Cao, J. Wang, and Y. Xia, "Combined electricity load-forecasting system based on weighted fuzzy time series and deep neural networks," *Engineering Applications of Artificial Intelligence,* vol. 132, 2024.
[32] X. Dong, Y. Luo, S. Yuan, Z. Tian, L. Zhang, X. Wu *et al.*, "Building electricity load forecasting based on spatiotemporal correlation and electricity consumption behavior information," *Applied Energy,* vol. 377, 2025.
[33] H. Xu, F. Hu, X. Liang, G. Zhao, and M. Abugunmi, "A framework for electricity load forecasting based on attention mechanism time series depthwise separable convolutional neural network," *Energy,* vol. 299, 2024.
[34] D. Jeong, C. Park, and Y. M. Ko, "Short-term electric load forecasting for buildings using logistic mixture vector autoregressive model with curve registration," *Applied Energy,* vol. 282, pp. 116249, 2021.