Deep Learning with Hybrid Attention for Power System Net Load Forecasting: A CNN-LSTM-GRU Approach

Yutong Zhou *

School of Electrical Engineering, Shanghai University of Electric Power, Shanghai, 200000, China

* Corresponding Author Email: zhouyutong@mail.shiep.edu.cn

Abstract. As the share of renewable energy in power systems continues to increase, net load forecasting has become significantly more challenging, with prediction errors under highly variable weather conditions often exceeding 300–600 MW using conventional methods. This paper introduces a hybrid attention model based on a Convolutional Neural Network (CNN) Long-Short-erm Memory (LSTM) -Gated Recurrent Units (GRU) architecture for net load forecasting. The model integrates a convolutional neural network for spatial feature extraction, long short-term memory and gated recurrent units for multi-scale temporal modeling, and an attention mechanism to dynamically weight relevant time steps. It processes raw net load data directly, avoiding the computational cost and inaccuracies associated with traditional decomposition techniques, which can introduce delays of several minutes and amplify errors during rapid changes. Validation under severe weather conditions shows that the proposed model achieves a mean absolute percentage error (MAPE) of 1.96% and a mean absolute error (MAE) of 42.85 MW, outperforming standalone LSTM, GRU, and CNN models by a margin of 18–32% in MAPE under similar conditions. This performance enables more reliable grid dispatch and contributes to maintaining system operational stability.

Keywords: Net Load Prediction, CNN-LSTM-GRU, Attention Mechanism.

1. Introduction

With more photovoltaic and wind renewable energy into the power grid, the net load forecasting has become increasingly complicated. Net load, which is the difference between the entire grid load and renewable energy generation, makes the forecasting accuracy of net load the straight-forward influence of dispatching effects of the grid, the precision of the power generation planning and the safety of the system. Under these circumstances, forecast net load faces with two major challenges [1]. Renewable energy's high volatility and uncertainty: the wind energy and photovoltaic output are heavily dependent on meteorological factors, and its volatility is fully and directly reflected in the net load, which results in an abrupt increase of the difficulty for forecasting. Poor quality and incomplete availability of data: historical load data might have outliers, and historical renewable energy output data in the remote area is still far from being complete [2]. Classical forecasting models such as autoregressive integrated moving average and exponential smoothing cannot deal with high-dimensional nonlinear correlations, and thus result in large forecast errors during peak load periods and price fluctuation periods. Deep learning models provide new approaches to net load forecasting, but single architecture models still face challenges in the aspects of capturing the subtle spatiotemporal attributes [3].

Recently, hybrid deep architectures display remarkable superiority in time series forecasting tasks. Convolutional neural networks (CNN) can effectively learn the local spatial features, and the recurrent neural networks such as Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU) are designed to grasp the long-term dependencies. Attention mechanism will enhance the model's ability to focus on the most important time steps. By taking advantages of the above two methods, the proposed hybrid CNN-LSTM-GRU model with an attention mechanism achieves a Mean Absolute Percentage Error (MAPE) of 1.96% and a Mean Absolute Error (MAE) of 42.85 MW on average over the test period. The model consistently outperforms individual benchmark models—including standalone LSTM, GRU, and CNN architectures.

2. Theoretical methods

2.1. Method for Obtaining Net Wind and Solar Power Generation Data

2.1.1. Wind power

The power output of wind turbines primarily depends on wind speed. The available power generated by wind passing through turbine blades can be mathematically expressed as a function of wind speed, as shown in the equation (1). Here, v represents wind speed, ρ_0 denotes air density, η_w indicates the turbines efficiency, and A denotes the swept area of the turbine blades. This area is defined by equation (1, 2), where D represents the diameter or length of the turbine blade [4].

$$P^{WT} = \frac{1}{2}\rho_0 A v^3 \eta_w \tag{1}$$

$$A = \frac{\pi}{4} D^2 \tag{2}$$

However, the actual output power of a wind turbine is limited by its rated operating wind speed [5]. Typically, wind turbines have specific start-up wind speedv_{ci}, shutdown wind speed v_{co}, and rated wind speed v_r. These parameters, combined with the current wind speed, determine the turbines' output power. The output power of a single wind turbine can be determined by the equation (3):

$$P^{WT}(v) = \begin{cases} \frac{1}{2}\rho_0 A v^3 \eta_w, & v_{ci} \le v < v_r \\ P_r, & v_r \le v \le v_{co} \\ 0, & v < v_{ci} \text{ or } v > v_{co} \end{cases} \tag{3}$$

Using this method, wind power output data can be calculated in each time series interval by combining wind speed information collected.

2.1.2. Photovoltaic power generation

The solar power generated by photovoltaic (PV) systems primarily depend on the solar irradiance reaching its surface. The output power is expressed by the equation (4), where η_{pv} represents the efficiency of the PV module, S_{pv} represents the modules surface area, and I_C represents the total solar irradiance received by the PV panel [6].

Nevertheless, the efficiency of photovoltaic systems generally exhibits non-constant characteristics and fluctuates in accordance with cell temperature, thereby exerting an influence on power output. In the existing literature, this process is commonly modeled through a linear regression model predicated on standard test conditions (STC), which integrates temperature-dependent correction factors [7]. The temperature-dependent relationship is contingent upon the material of the solar panel and is represented by equation (5): where $I_{C,ref}$ and T_{ref} represent the reference irradiance and cell temperature under STC conditions, $P_{pv,ref}$ denotes the rated output power at STC, γ_{ref} is the temperature coefficient, and T_{cell} is the actual cell temperature.

$$P_{pv} = \eta_{pv} S_{pv} I_C \tag{4}$$

$$P_{pv} = \frac{I_C}{I_{C,ref}} P_{pv,ref} \cdot \left[1 - \gamma_{ref}\left(T_{cell} - T_{ref}\right)\right] \tag{5}$$

Equation (5) effectively delineates the correlation between battery temperature and output power. Nevertheless, prior to utilizing this equation, it is imperative to ascertain the battery temperature, a value contingent upon environmental conditions. The energy transfer among diverse media is instigated by temperature differentials and typically transpires via three heat transfer modalities: heat absorption, heat radiation, and heat convection [8].

Within a closed system, the three heat transfer processes will attain thermal equilibrium. Consequently, under steady-state circumstances, the thermal balance of a photovoltaic system can be represented by equation (6). Where q_s, q_r, and q_c represent heat absorption, heat radiation, and heat convection rates, respectively, and P_{pv} denotes the photovoltaic power output.

$$q_s + q_c + q_r - P_{pv} = 0 \tag{6}$$

In PV panel systems, thermal absorption is defined by equation (7), where α_{cell} represents the absorption rate of the photovoltaic cell. Thermal radiation is given by equation (8), where σ is the Stefan-Boltzmann constant, ε_a is the emissivity of the environment, T_a is the ambient temperature, and ε_{cell} is the emissivity of the cell. Thermal convection is defined by the formula (9), where h_c is the convective heat transfer coefficient.

$$q_s = \alpha_{cell} I_C S_{pv} \tag{7}$$

$$q_r = S_{pv}\sigma[\varepsilon_a T_a^4 - \varepsilon_{cell} T_{cell}^4] \tag{8}$$

$$q_c = -h_c S_{pv}(T_{cell} - T_a) \tag{9}$$

The convective heat transfer coefficient, defined by the equation (9), $h_c$is the sum of the natural convection coefficient $h_{c,free}$ and the forced convection coefficient $h_{c,forced}$.These values depend on environmental conditions, particularly wind speed and temperature. For photovoltaic panel systems, these coefficients are defined by the equation (10-12).

$$h_c = h_{c,free} + h_{c,forced} \tag{10}$$

$$h_{c,free} = \frac{0.1k_0}{L}\left(\frac{g\rho_0\beta_0 C_{p0}}{\mu_0 k_0}\right)^{\frac{1}{3}}(T_{cell} - T_a)^{\frac{1}{3}} \tag{11}$$

$$h_{c,forced} = \begin{cases} \frac{0.664k_0}{L}\left(\frac{\rho_0 v}{L}\right)^{\frac{1}{2}}, & v < 3.3037_{\mathrm{m/s}} \\ \frac{0.037k_0}{L^{10}\mu_0^{15}}(\rho_0 v)^{\frac{4}{5}}, & v > 3.3037_{\mathrm{m/s}} \end{cases} \tag{12}$$

This method can determine the battery temperature depending on PV irradiance, air temperature and wind speed, the parameters are obtainable from measurement from 2015 – 2018. And with such a functional model the temperature of PV cell can be calculated for one time period, and then the calculation of power output data of PV system becomes possible.

2.2. Data Preprocessing

In this study, an eight-year hourly load dataset is adopted and reconstructed with sliding window setting. By setting historical sample size and prediction horizon, long training samples reflect both seasonality and daily variation and other time-varying effects from the data. Data standardization is conducted using the minimum-maximum normalization equation (13), transforming the original data into the [0,1] range [9]. In view of the characteristics of time series data, the input-output mapping is constructed by sliding step by step to provide high diversity of training samples for model learning.

$$x_{norm} = \frac{x - x_{min}}{xmin_{max}} \tag{13}$$

2.3. CNN-LSTM-GRU Model

Conventional CNN-LSTM models exhibit three inherent disadvantages to net load forecasting. First, since LSTM works as a stand-alone model, its ability to model the temporal dynamics is limited in that it is hard to represent various multi-scale temporal patterns ranging from short-term patterns to long-term trends. Second, the information utilizing mechanism is rather naive, since it does not have an attention mechanism to dynamically focus on the important time-steps necessary for the prediction. Therefore, the important signals will be overwhelmed by noises. Third, efficiency and robustness are also inadequate. It is worth noting that just adding the complexity of LSTM only for the purpose of enhancing the performance tends to cause explosion of parameters, overfitting, and decrease the generalization performance on the complex and dynamic scenario. Therefore, to address

these problems, in this paper present an effective hybrid architecture, which contains multi-scale GRU and dynamic attention mechanism as illustrated in the next workflow Fig.1[10].

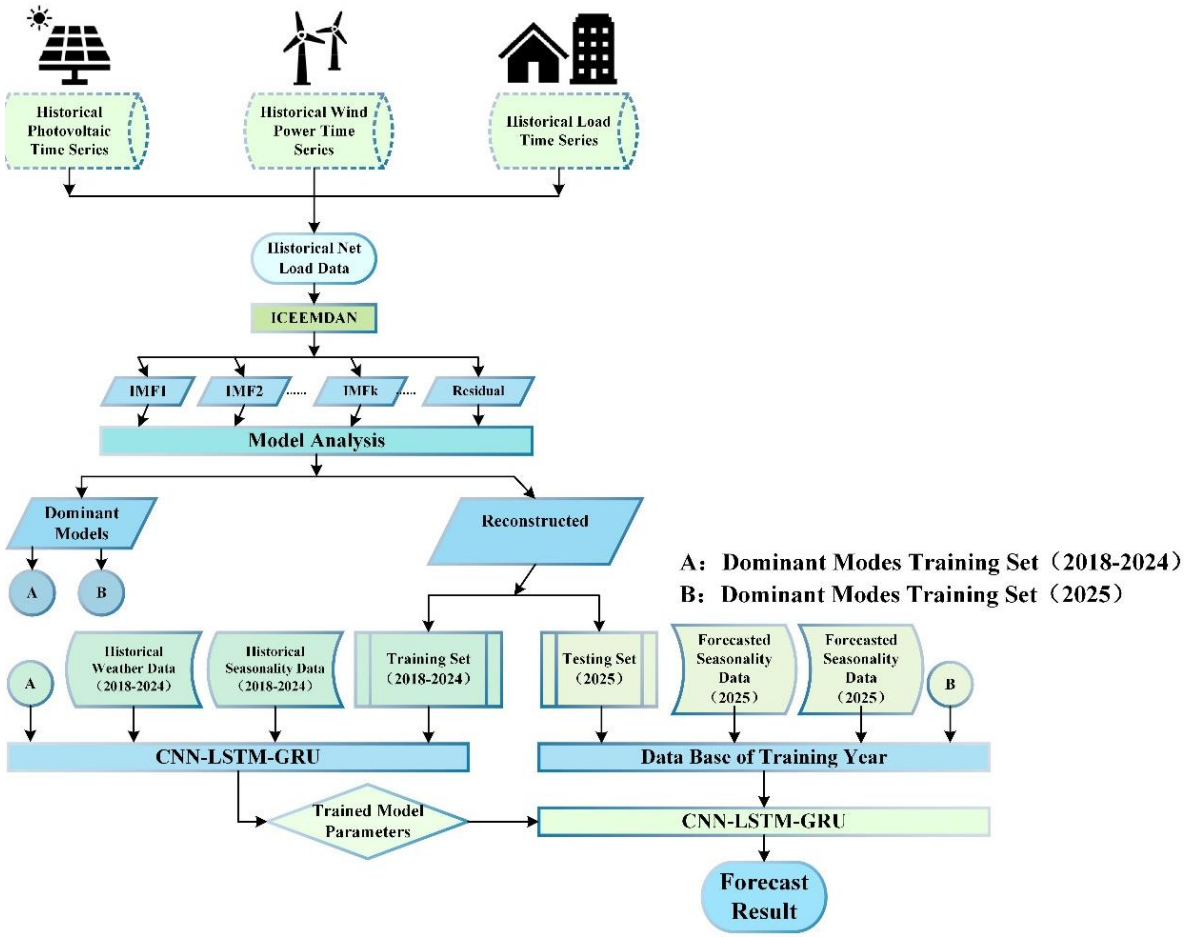

Figure 1. CNN-LSTM-GRU Model (Picture credit: Original)

2.3.1. CNN module

This module contains two successive one-dimensional convolutional layers. The first convolutional layer is a 128 5×5 convolution kernels with a large receptive field, used to capture the macro trends and cycle pattern in the sequences and followed by a ReLU activation function to introduce non-linear transform; and the second convolutional layer is a 128 3×3 convolution kernels, with a small receptive field, used to adjust the capture local fluctuation, fine details [11]. Detailed workflow is shown in Fig.2 as follows.

Each Convolutional layer has a one dimension (1D) max pooling layer that down samples the sequence length, minimizing the computational complexity while increasing the spatial invariance of features and therefore helping towards a model's robustness [12]. In addition, each convolutional block will have a set of batch normalizations and Dropout layers. Batch normalizations fix the activation distributions to speed up the training. Dropout is to avoid overfitting by randomly dropping some connections of neural units.

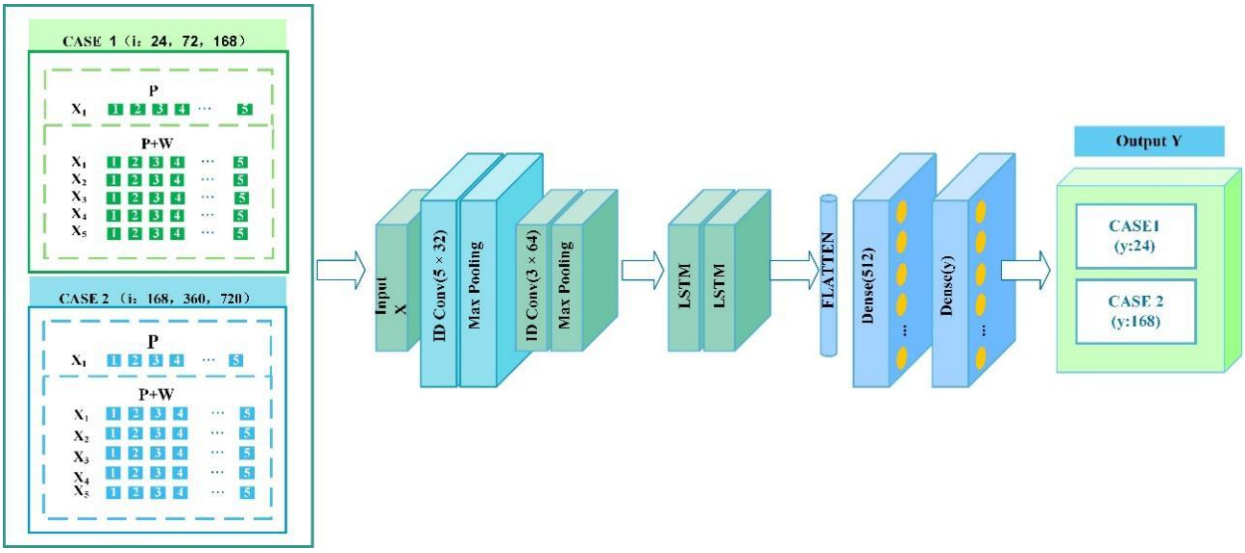

Figure 2. Convolutional Layer (Picture credit: Original)

Since deep model suffers from the gradient vanishing problem, the module explicitly incorporates a residual connection from the second convolution layer to the first convolution layer. And adjust the dimensionality of the short connection by a 1×1 convolution kernel so that the output of the main path can be added to the one from the second layer [13]. This allows the model learns to operate in both residual and original mapping simultaneously and benefits the training dynamic. The output of this module is a high-order feature map, which converts the original input sequence into a feature sequence rich in local semantic information, laying a solid foundation for subsequent time series modeling.

2.3.2. LSTM module

The feature sequences extracted by the CNN module are fed into a carefully designed hybrid recurrent neural network module to capture the complex long-term and short-term temporal dynamics in the data.

Bidirectional LSTM Layer: The architecture begins with a bidirectional LSTM layer, each direction containing 128 hidden units. Leveraging its meticulously designed gating mechanisms (input gate, forget gate, output gate), the LSTM effectively captures and retains temporal dependencies across extended periods. The bidirectional architecture enables the model to analyze sequences both forward and backward, utilizing past and future contextual information for predictions [14]. This dual-directional capability is crucial for understanding cyclical patterns such as daily and weekly rhythms.

Bidirectional GRU Layer: the module follows the bidirectional GRU layer with 64 hidden units in each direction. GRU is another variant of LSTM that groups together the input and forget gates into a single update gate, hence, achieves similar performance as LSTM, but with fewer parameters and better computational efficiency [15]. Here, the GRU layer is used to capture the short-term temporal dynamics possibly missed by LSTM.

Hybrid Structure: the module adopts the hybrid structure of LSTM-GRU for functional complementarity. LSTM can effectively model long-term temporal dependencies. GRU can capture medium-term or short-term temporal dependencies with light weight and efficiency. Together they make strong time-series feature encoder. The module outputs a sequence of hidden states with rich temporal dependencies information.

3. Results

Fig. 3 and Fig.4 show a comparison of the performance of different models on a clear but cloudy day with variable net load patterns on December 31, 2024.

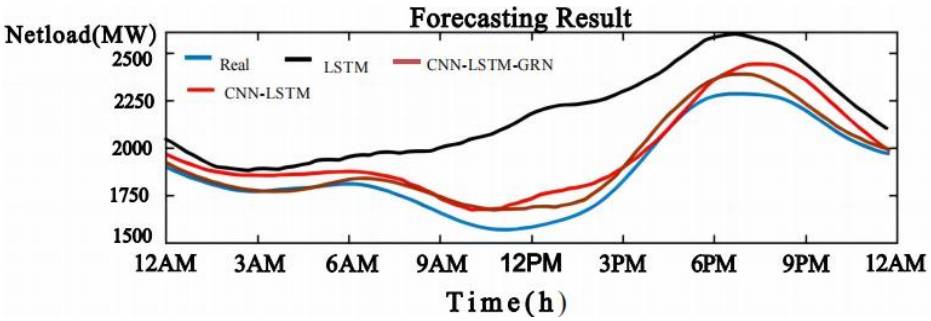

Figure 3. Forecasting Result (Picture credit: Original)

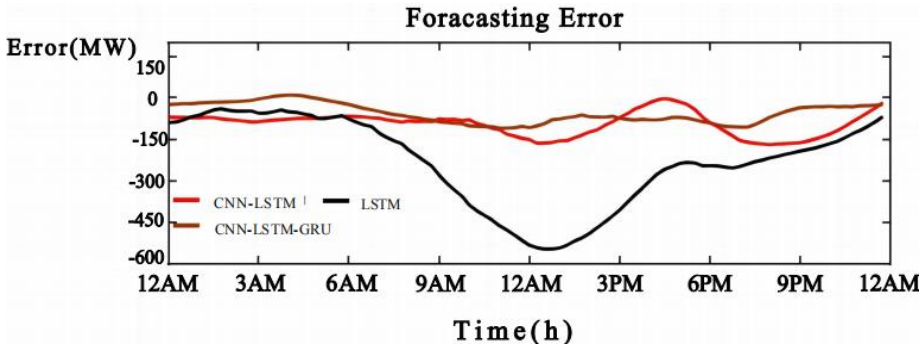

Figure 4. Forecasting Error (Picture credit: Original)

The experiment was based on a desktop computer with a 10th Gen Intel Core i9 processor and 16 GB of RAM. Table 1 shows the comprehensive performance of CNN-LSTM-GRU model in the prediction of daily net load in 2019.

As observed from Fig.3, Fig.4 and Table 1, the figures offer a clear visual demonstration of the superior performance of the CNN-LSTM-GRU model. In Fig.3, the brown curve representing the CNN-LSTM-GRU model adheres more closely to the actual net load trajectory throughout the day, particularly during the dynamic morning ramp-up and the evening peak periods, indicating its exceptional ability to track complex, real-time changes. This visual superiority is quantitatively explained by the Fig.4, which shows that the CNN-LSTM-GRU model maintains the smallest and most stable error margin, consistently avoiding the large error spikes of 300–600 MW that plague the other models, especially during high-volatility hours. This performance aligns perfectly with the paper's quantitative results, which report a low average MAE of 42.85 MW and a MAPE of just 1.96%. This consistent accuracy stems from the model's hybrid architecture, where the CNN effectively extracts local spatial patterns, the LSTM and GRU units collaboratively capture both long-term and short-term temporal dependencies, and the attention mechanism dynamically prioritizes the most critical time steps. Consequently, the model demonstrates a robust capacity to handle the nonlinear and non-stationary characteristics of net load data under diverse weather conditions, solidifying its advantage over simpler benchmark models like standalone LSTM or CNN-LSTM.

Table 1. The Values of MAE and MAPE

Time	CNN-LSTM-GRU	
Date	MAPE (%)	MAE(MW)
01/01-01/14	1.5	31.51
01/15-01/28	1.88	4,109
01/29-02/11	2.09	479
02/12-02/25	1.44	32.02
02/26-03/11	2.7	60.55
03/12-03/25	1.88	40.85
03/26-04/08	1.66	34.1
04/09-04/22	1.42	28.62
04/23-05/06	1.67	35.02
05/07-05/20	1.44	31.88
05/21-06/03	1.27	26.72
06/04-06/17	1.87	44.41
06/18-07/01	2.06	49.11
07/02-07/15	1.92	42.6
07/16-07/29	1.67	37.64
07/30-08/12	1.39	31.53
08/13-08/26	1.56	35.29
08/27-09/09	1.86	40.22
09/10-09/23	2.02	40.5
09/24-10/07	1.72	32.83
10/08-10/21	1.88	35.98
10/22-11/04	2.15	42.15
11/05-11/18	3.64	78.91
11/19-12/02	2.21	43.25
12/03-12/16	3.52	96.13
12/17-12/31	2.47	53.17
Average	1.96	42.85

The real value of the load (the blue curve) and the predicted values according to the model (the red curve) for the train set are shown in Fig.5 (a). It is possible to have a qualitative inspection on how good the fit of the model to the train data by is comparing these values. The real values for the load (blue curve) and the values forecasted by the model (red curve) for the test set shown in Fig. 5 (b) are used to examine the potential generalization of the model to an unseen input. The comparisons of the performance in the test and train sets can provide some insights about any overfitting. The errors reported in the title represent the model's real performance. The histogram shown in Fig.5 (c) represents the distribution of the prediction errors (i.e. actual value - predicted values). The distribution shape of the error (symmetry and where the peak(s) are, etc.) tells us how concentrated or how scattered the errors are in the training set. The mean error reported in the title serves as the mean value of the error distribution. For the test set, see Fig.5 (d) for the error distribution of the prediction. Compared to the training set, this method could make a guess as to the error characteristics of the model for both sets of data.

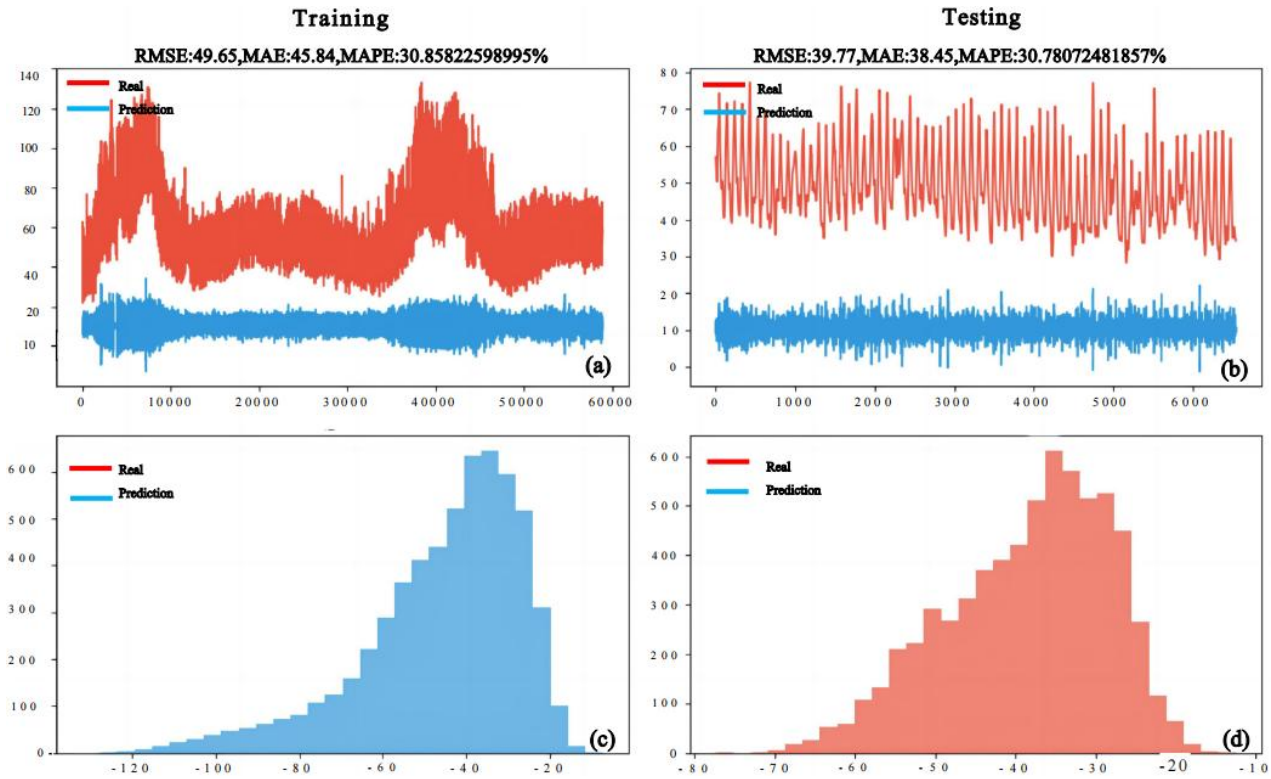

Figure 5. Models' prediction accuracy (Picture credit: Original)

4. Conclusion

This study successfully developed and trained a hybrid CNN-LSTM-GRU model with an attention mechanism for power system net load forecasting—a task growing in complexity due to the increasing integration of renewable energy sources. The proposed model demonstrates significant superiority over conventional forecasting systems by leveraging the spatial feature extraction capability of CNN, the high-order temporal modeling strengths of LSTM and GRU networks, and the selective focus enabled by the attention mechanism. Experimental results confirm the model's enhanced performance, achieving the MAPE of 1.96% and the MAE of 42.85 MW on average during the test period, significantly outperforming individual LSTM, GRU, and CNN models. The hybrid architecture proves effective in capturing complex spatiotemporal correlations in net load data under diverse weather conditions, offering valuable technical support for power grid dispatch and stable system operation.

Even with some deficiencies, this work illustrates the promising capacity of a hybrid approach of deep learning models for net load forecasting in today's power systems. For the coming research, model tuning to enhance computational performance, addition of probabilistic forecasting, and implementation of more effective methodologies for interpretability are planned, in order to make them appeal to a broader industrial audience.

Nevertheless, there are some limitations of the proposed methodology that needs to be addressed. The model structure adds high computational complexity, demanding extensive training times and memory footprint compared with other lightweight models. The model structure hybridity, which is beneficial to achieve a better performance, requires proper parameter calibration and potentially present difficulties for on-line implementation on a resource constrained scenario. Second, the performance of the model relies on the quality and availability of the data and is thus subject to the general weakness of data driven models, the models do not yield meaningful results for incomplete or noisy input. Third, while the attention mechanism increases the interpretability of the model to certain degree, it cannot completely dispel the "black box" character of deep learning models that

hinders the practical deployment of this approach in applications requiring full operational transparency.

References

[1] Van der Meer, D. W, Munkhammar, J. & Widén, J. Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals. Solar Energy, 2018, 171: 397 – 413.

[2] Tziolis, G, Lopez-Lorente, J, Baka, M.-I., Koumis, et al. Direct short-term net load forecasting in renewable integrated microgrids using machine learning: A comparative assessment. Sustainable Energy, Grids and Networks, 2024.

[3] Sreenu, S, Sharma, K. C, Bhakar, R. Gumbel copula based aggregated net load forecasting for modern power systems. IET Generation, Transmission & Distribution, 2018, 12: 4348 – 4358.

[4] A. Wahab, M.A. Tahir, N. Iqbal, A. Ul-Hasan, F. Shafait, S.M. Raza Kazmi. A novel technique for short-term load forecasting using sequential models and feature engineering. IEEE Access, 2021, 9: 96221 - 96232.

[5] Zhang Yun, Zhou Quan, Caixin Sun, Shaolan Lei, Yuming Liu, and Song Yang. RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Transactions on Power Systems, 2002, 23 (3): 853 - 858.

[6] T.Y. Kim, and S.B. Cho. Predicting residential energy consumption using CNN-LSTM neural networks. Energy, 2019, 182: 72 - 81.

[7] Monowar Hossain, and Saad Mekhilef. Application of Extreme Learning Machine for short term output power forecasting of three grid-connected PV systems. J Cleaner Prod, 2017, 167: 395 - 405.

[8] Liu, C, Zhang, K, Xiong, H, Jiang, G, Yang, Q. Temporal self-attention network for medical concept embedding. IEEE International Conference on Data Mining, 2020.

[9] Md. Rayhan Ahmed, Salekul Islam, A.K.M. Muzahidul Islam, Swakkhar Shatabda. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Systems with Applications, 2023.

[10] Himmet ÖZARSLAN, Ihsan ULUOCAK. CNN-LSTM and CNN-GRU based estimation of NOx conversion efficiency in diesel engine exhaust system. Fuel, 2026.

[11] K.e. Yan, X. Wang, Y. Du, N. Jin, H. Huang, H. Zhou. Multi-Step Short-Term Power Consumption Forecasting with a Hybrid Deep Learning Strategy. Energies 11, 2018.

[12] Md. A. Istiake Sunny, M. M. S. Maswood, et al. Deep Learning Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 2nd Novel Intelligent and Leading Emerging Sciences Conference, 2020.

[13] Osaka Rubasinghe, Tingze Zhang, Xinan Zhang, San Shing Choi, Tat Kei Chau, Yau Chow, Tyrone Fernando, Herbert Ho-Ching Iu. Highly accurate peak and valley prediction short-term net load forecasting approach based on decomposition for power systems with high PV penetration. Applied Energy, 2023.

[14] P. Ravindran, A. Costa, R. Soares, and A.C. Wiedenhoeft. Classification of CITES-listed and other neotropical Meliaceae wood images using convolutional neural networks. Plant Methods, 2018.

[15] Vasileios Pentsos, Spyros Tra goudas, Senior Member, IEEE, Jason Wibbenmeyer, and Nasser Khdeer, A Hybrid LSTM-Transformer Model for Power Load Forecasting. IEEE Transactions on Smart Grid, 2025.