Electrical Power and Energy Systems 158 (2024) 109975

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

ATTnet: An explainable gated recurrent unit neural network for high frequency electricity price forecasting

Haolin Yang [a], Kristen R. Schell [b,a,*]

[a] Industrial & Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[b] Mechanical & Aerospace Engineering, Carleton University, Ottawa, K1S 5B6, Canada

ARTICLE INFO

Keywords:
Deep learning
RNN
Time series analysis
Electricity price forecasting
Attention mechanism
SHAP
NYISO
Real-time market

ABSTRACT

The primary contribution of this study is the proposal of an explainable deep-learning neural network (ATTnet) that employs an attention mechanism to achieve accurate electricity spot price forecasting and an explainable model pipeline. The concise, single-stream network consists of a 5-head attention mechanism and gated recurrent units, which have been developed to model the temporal dependencies of the volatile market data. In addition to introducing a novel neural network architecture for volatile time series data, this study makes a substantial contribution by investigating prediction factors in two ways: temporally via the attention scores from the input sequences and globally via feature Shapely values. In real-time electricity price prediction, historical prices, temperature, hour, and zonal load are found to be the most important variables. The deep learning model was tested on real-time price profiles from eight generators within the New York Independent System Operator (NYISO) network. The proposed model achieves performance gains of 21% in MAE and 22% in MAPE over the state-of-the-art benchmark methods.

1. Introduction

Electricity price forecasting can be achieved through reliance on historical prices alone, such as in traditional time series modeling like ARIMA [1,2] and LASSO [3]. Electricity prices are known, however, to be influenced by more fundamental factors, such as the weather, which can cause an increase in demand leading to an increase in price [4]. Several studies have included exogenous variables in an attempt to more accurately forecast price. System load, or demand, is often included [5,6], as this market variable is directly related to the price. Weather variables, such as temperature [7], renewable power supply [8], wind speed and humidity [9], have been shown to improve forecasting accuracy. Including time factors, such as hour, week and month [10] have also proven useful. Other fundamental factors affect electricity prices, such as fuel costs, available supply based on scheduled or forced maintenance, and the available reserve capacity within the system.

Given such a large pool of possible input data, modelers typically incorporate variables based on their expert knowledge of the market operations, as well as data availability constraints. This is because an optimal, empirical method for input variable selection does not exist. Deep learning (DL) models and parallel computation methods offer the capacity to include an expanded number of input variables to help improve model predictive accuracy. However, very few studies in the DL field of price forecasting [11] have analyzed how to interpret these fundamental factors' affect on price. New methods for interpreting prediction drivers have been applied to other energy-related prediction tasks.

1.1. Interpreting prediction drivers in deep learning models

Deep learning models have the potential to play an essential role in multiple energy-related tasks such as planning, grid operations, and decision-making. However, unfortunately, most of them lack transparency [12]. There is a growing field of research to enhance the interpretability of black-box deep learning models, to provide a better understanding of forecasts and predictions to decision-makers and grid operators. To date, Dong et al. [13] have developed a generative adversarial network (GAN) as an interpretability technique using adjustable feature vectors. Moreover, Zhang et al. [14] have designed an interpretable preventive control model for power system transient security based on XGBoost and LIME. Due to their flexibility and applicability to different types of the DL architecture, two other methods for enhancing interpretability have also recently been proposed — the attention mechanism [15] and Shapley values [16]. Both methods are applied here to electricity price forecasting (EPF).

[*] Corresponding author.
E-mail address: kristen.schell@carleton.ca (K.R. Schell).

https://doi.org/10.1016/j.ijepes.2024.109975
Received 27 July 2023; Received in revised form 6 March 2024; Accepted 30 March 2024
Available online 6 April 2024

1.1.1. Attention mechanism

As a state-of-the-art deep learning algorithm, the attention mechanism is able to capture information expressed temporally within the positions of a sequence [17]. The attention mechanism has been widely used in multiple fields such as: geo-spatial analysis [18,19], anomaly detection [20], time series prediction [21,22], traffic flow forecasting [23], climate pattern recognition [24] and financial management [25]. Within the areas of energy and power systems, there have similarly been several applications. In 2019, a BiLSTM hybrid model [26] with attention was proposed for power load forecasting, and, in 2020, [27] a self-attention convolutional neural network (CNN) was established to detect transmission line faults. Furthermore, Heidari and Khovalyg proposed an attention-based LSTM for energy demand prediction, which was designed for small, solar-assisted water heating systems [28]. Combined with Gated Recurrent Units (GRU), Niu et al. [29] proposed an attention-based framework for wind power prediction. In 2021, Zhang et al. [30] proposed a CNN for wind speed prediction. In 2022, the sequence-to-sequence building energy prediction task is achieved via an encoder–decoder structure based on LSTM and attention [31]. In the meantime, a dual-state-attention-based LSTM was proposed for short-term load forecasting [32], while an attention-LSTM model [33] was designed for reactive and active load prediction at five substations. Except for the deep learning model proposed by this study with multiple-head attention modules, the attention mechanism has not yet been applied to the real-time electricity price forecasting problem, as far as the authors are aware.

1.1.2. Assessing feature importance via Shapley values

As applied to feature importance assessment, a method based on Shapley values (SHAP) calculates the marginal contribution of all combinations of features towards the prediction [34]. The methodology, developed via theoretical game theory, has been used in power system research areas such as emergency control [35,36], event prediction [37], load forecasting [38,39] and power plant diagnostics [40]. While SHAP has been applied to electricity price forecasting, to the best of the authors' knowledge, no studies have yet applied SHAP to the volatile real-time market. In 2021, Li and Becker proposed an LSTM autoencoder with kernel SHAP feature explanations [41]. In 2022, Tschora et al. [11] tested different machine learning models and used SHAP to conduct the feature analysis, with case studies in the French, German, and Belgian day-ahead markets.

1.2. Contributions

To the best of the authors' knowledge, a deep learning neural network with a multi-head attention mechanism has not yet been developed for the problem of real-time electricity price forecasting. Such a model is hypothesized to have the ability to deal with the noise and volatility of real-time power pricing data. However, this hypothesis has not yet been tested. In addition to developing a new model, this work also studies methods to identify the most important features impacting price forecasting. The following summarizes the principal contributions of this work:

- Development of an innovative, simpler and more accurate deep learning architecture for highly volatile, real-time electricity price forecasting.
- Development of a novel, multi-head attention mechanism (five-head) embedded in the architecture, which skillfully learns the temporal similarity score to model high-frequency dependencies in the spot price time series.
- Analysis and visualization of feature importance in the DL model via the SHAP method and attention scores.

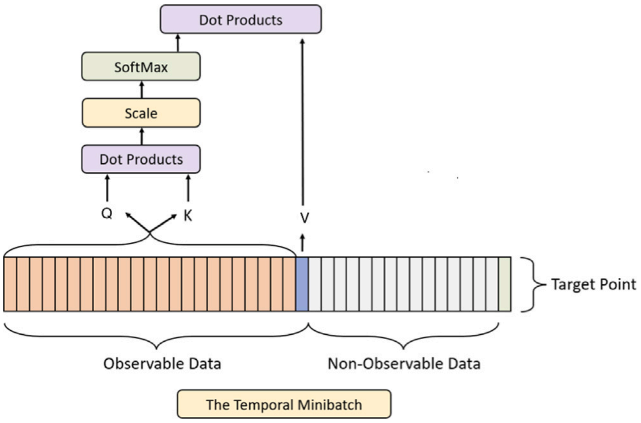

Fig. 1. Multi-head self-attention.

2. Methodology

2.1. Multi-head self-attention

The attention mechanism gives important features higher weight from different dimensions [42,43]. Such 'highlighting' can help the DL model capture informative and relevant components, even in a noisy sequence of input data, such as real-time electricity prices. This is mostly due to the fact that multi-head attention, which was analyzed further here in Section 4.3 (see Fig. 8), will detect the influence of different temporal patterns on the dynamics of electricity prices. Specifically, multi-head attention based on the Scaled Dot-Product Attention (SDA) equation [17] is applied in this study. The formulation of the SDA is as follows:

$$\text{Attention } (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where matrix Q is the packed queries, and matrices K and V are Keys and Values. The dot-product between Q and K is divided by the scaling factor $\sqrt{d_k}$, the square-root of the dimensionality of the K matrix, to keep the gradients stable and prevent a vanishing gradient problem. The softmax function is applied to evaluate the probabilistic weights of the values, V. Unlike when the SDA is used in a translation task, in EPF the target sequence cannot be directly put into the attention mechanism, as it occurs in the future. Moreover, because real-time market bidding must be closed at least 75 min before dispatch [44], there is also a prediction gap ahead of the target that must be enforced, in which temporal dependencies are not observable. As Fig. 1 shows, the input of values, V, are not the traditional target sequence, but here are the last temporal features of the input sequence. This estimates the target point as it is closest to the temporal position of targets, save the un-observable data gap that must be enforced. To model information from different subspace representations [17], the SDA is computed multiple times in parallel, and the independent outputs are concatenated and then transferred to the subsequent layers of the DL model. Regardless of the sequential dependencies, the multi-head attention layer only processes the input data as a set of arrays. Therefore, a positional encoding vector is added to the input vectors, to provide the relative position information of the dependencies in the temporal sequence [17].

2.2. Deep learning architecture

Instead of a multi-branched architecture, which has been previously proposed for electricity price forecasting [45,46], the authors have designed a vertical neural network without subdivisions, shown in Fig. 1, inspired by a transformer encoder architecture. However, due to

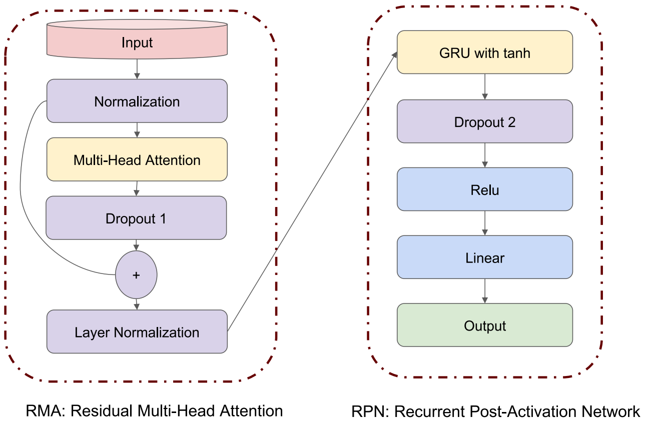

Fig. 2. Schematic of ATTnet deep learning architecture.

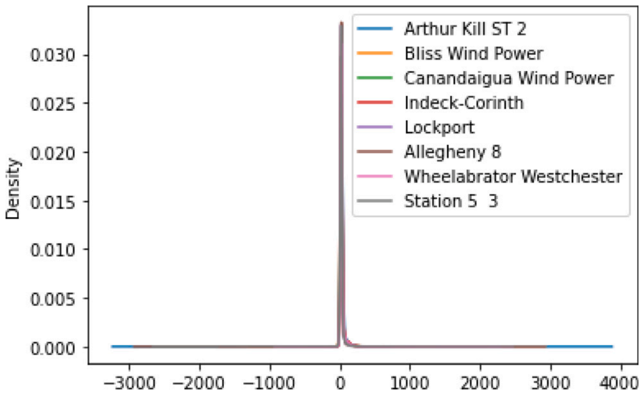

Fig. 3. Density plot showing real-time price [$/MWh] variability of the selected generators.

the high volatility of real-time electricity price data, the conventional transformer encoder architecture cannot guarantee a robust output for this application. Therefore, the traditional model has been modified and upgraded by the authors. The proposed model is divided into two main parts with differing functions: (1) the Residual Multi-head Attention (RMA) and (2) a Recurrent Post-Activation Network (RPN), as shown in Fig. 2. Within RMA, the input is normalized and then fed into the multiple head attention. The normalization layer is applied to transform the input data into scales for better training and performance. The output of the Multi-head Attention mechanism is added back to the normalized input and then transferred to a dropout layer, followed by a normalization layer. The dropout technique is a regularization strategy employed to mitigate over-fitting by randomly deactivating some nodes within a layer during the training process, hence promoting independence among the units [47]. Such residual connection has been proven to facilitate model training and prevent the vanishing gradient problem [48]. Layer normalization is applied to the output from the previous layer in a batch independently, which enhances the training speed [49]. The processed tensors are then input to a GRU [50] followed by a ReLU activation function to capture the recurrent temporal dependencies. The rationale behind choosing the GRU as the foundation of the RPN lies in its ability to achieve comparable or superior performance while consuming fewer computational resources with less parameters, in contrast to other types of neural networks like LSTM [51, 52]. The ReLU is added to handle the nonlinearity of the model, while the subsequent linear layer is used to reshape the tensor into the desired shape for prediction output. According to 24 sets of trials (i.e. eight generators with three unobservable window time frames), the proposed architecture improved performance by 24.67% in MAE and 13.32% in sMAPE compared to the original transformer encoder structure, and by 11.14% in MAE and 7.70% in sMAPE compared with the Informer encoder (another version of transformer-subclass model based on a convolutional neural network [53]). More information regarding the comparison between the proposed model and state-of-the-art models can be found in Section 4.

2.3. Data processing and sampling

A down-sampling method to different price quantiles is used to build up a robust training dataset which minimizes the effect of the long-tailed distribution of volatile real-time electricity price.

2.3.1. Sequence-to-point pair generation

The forecasting task is formulated as the sequence-to-point problem, which means a minibatch of the temporal information is used to predict a single target price in the future. Fig. 4 illustrates how input data sequences are created via two moving window methods. Firstly, the time series is transformed into batches based on a rolling sliding window and a day interval time step; these sampling methods are the same process utilized in other published EPF deep learning models [45,46].

Given a certain temporal input such as temperature or historical electricity price, $X(t) = \{x_1, x_2, x_3, x_4, \dots, x_n\}$, the rolling sliding window sequence-to-point method is formulated as follows:

$$ind = t - o - u - 1 \tag{2}$$

$$\hat{y}_t = f(x_{ind}, x_{ind+1}, x_{ind+2}, x_{ind+3}, \dots, x_{ind+o}) \tag{3}$$

where o represents temporal points that can be observed while u are the temporal points that cannot be observed (see Fig. 1; this is also referred to as the 'GAP' in Section 4). Due to market bidding constraints, some temporal points before the predicted target are not observable. These are modeled as an unobservable data gap, defined as the number of data points that cannot be observed before the predicted price point. ind represents the first temporal stamps among the rolling window sequence. The target is the tth output, $\hat{y}_t$.

The day interval time method formulates the sequence-to-point problem as follows:

$$ind = t - o - u - 1 \tag{4}$$

$$sp_k = x_{ind+o-V\times(L-n)}, \;\; k = 0, 1, 2, 3 \dots L \tag{5}$$

$$\hat{y}_t = f(sp_0, sp_1 \dots, sp_L) \tag{6}$$

where L is the series length and V is the temporal interval length between real-time electricity price. Eq. (5), sp_k is the symbol of one sampling point. The n is the temporal index of the sampling points.

Using the two sequence generation methods to process the features with aligned timestamps, the dimensions of one minibatch are $[o, 2\times d]$, where d is the number of features. To ensure the minibatches generated from the two methods can be concatenated and input into the single branch model, in this work o is equal to 288.

2.3.2. Data sub-sampling based on real-time electricity price quantiles

The steps of sub-sampling are as follows:

- Collect training data including features, minibatch and targets pairs $< X, y >$
- Separate the targets Y in the training data into bins of equal quantile size based on rank order. In this work, the targets are discretized into 10 quantile bins.
- The Evenly sub-sample the data from each normal range bins.

To avoid data leakage, this sub-sampling process is only conducted in the training data (see Fig. 5).

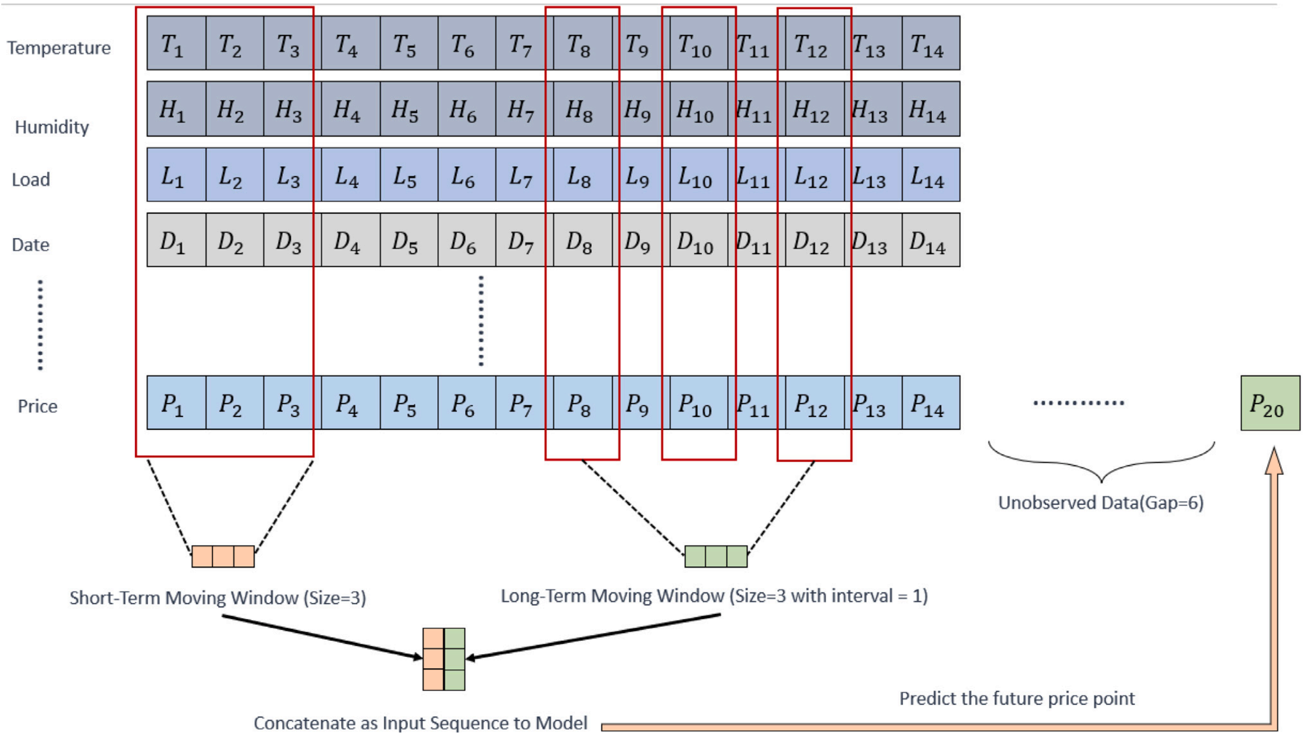
Temperature
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} T_{11} T_{12} T_{13} T_{14}
Humidity
H_1 H_2 H_3 H_4 H_5 H_6 H_7 H_8 H_9 H_{10} H_{11} H_{12} H_{13} H_{14}
Load
L_1 L_2 L_3 L_4 L_5 L_6 L_7 L_8 L_9 L_{10} L_{11} L_{12} L_{13} L_{14}
Date
D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10} D_{11} D_{12} D_{13} D_{14}
Price
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P_8 P_9 P_{10} P_{11} P_{12} P_{13} P_{14}
P_{20}
Unobserved Data(Gap=6)
Short-Term Moving Window (Size=3)
Long-Term Moving Window (Size=3 with interval = 1)
Predict the future price point
Concatenate as Input Sequence to Model

Fig. 4. Illustration of input data sampling method. The long-term moving window interval (here demonstrated as 1) changes by experiment, equal to the prediction interval gap that is enforced (see Section 4.2).

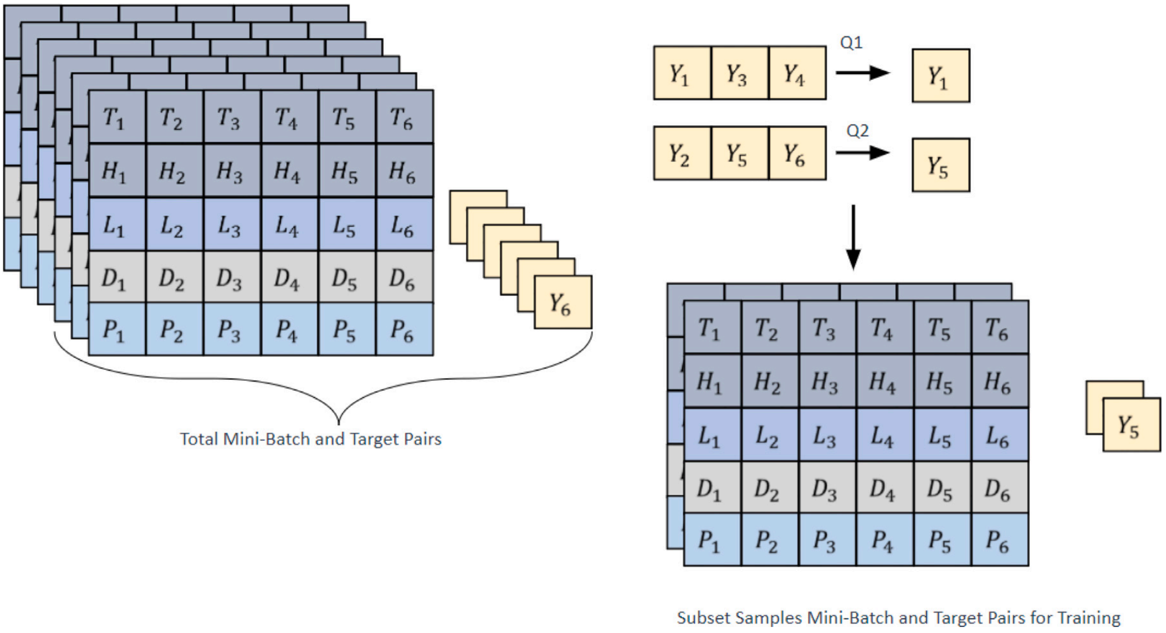
T_1 T_2 T_3 T_4 T_5 T_6
H_1 H_2 H_3 H_4 H_5 H_6
L_1 L_2 L_3 L_4 L_5 L_6
D_1 D_2 D_3 D_4 D_5 D_6
P_1 P_2 P_3 P_4 P_5 P_6
Y_6
Total Mini-Batch and Target Pairs
Y_1 Y_3 Y_4
Q1
Y_1
Y_2 Y_5 Y_6
Q2
Y_5
T_1 T_2 T_3 T_4 T_5 T_6
H_1 H_2 H_3 H_4 H_5 H_6
L_1 L_2 L_3 L_4 L_5 L_6
D_1 D_2 D_3 D_4 D_5 D_6
P_1 P_2 P_3 P_4 P_5 P_6
Y_5
Subset Samples Mini-Batch and Target Pairs for Training

Fig. 5. Schematic of mini-batch sub-sampling as quantiles of target distribution.

2.4. SHAP feature interpretation

In addition to the attention score, the SHapley Additive exPlanations (SHAP) [54–56] method is utilized to improve model explainability. As a visualization tool [57], SHAP is applied to explain the prediction of the proposed model by calculating the contribution of each feature towards the target output via its Shapley value [58]. As an additive method, SHAP measures the feature importance by calculating the marginal contribution on a subset of features. The effect value of each feature ϕ_i [59], is calculated as the weighted average of the difference in prediction between the model trained with the feature ($f_{(S\cup\{i\})}$) and without (f_s):

$$\phi_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}\left[f_{(S\cup\{i\})}(x_{(S\cup\{i\})}) - f_s(x_S)\right] \tag{7}$$

Where S is the subset of feature groups while N is the whole set of all features. However, it is challenging to exactly compute such values directly because of its intractable nature [54]. In this paper, the authors utilize the SHAP implementation called Gradient Explainer [60], which is an extension of integrated gradients [61]. Shapley values are

Table 1
Generator information.

Index	Name	PTID	Zone	Town	In-service date (month/day/year)	Type
0	Arthur Kill ST 2	23 512	J	Staten Island	8/1/1959	Steam Turbine
1	Bliss Wind Power	323 608	A	Bliss	3/20/2008	Wind Farm
2	Canandaigua Wind Power	323 617	C	Avoca	12/5/2008	Wind Farm
3	Indeck-Corinth	23 802	F	Corinth	7/1/1995	Combined Cycle
4	Lockport	23 791	A	Lockport	7/1/1992	Combined Cycle
5	Allegheny 8	23 528	C	Kittanning PA	10/1/1990	Conventional Hydro
6	Wheelabrator Westchester	23 653	H	Peekskill	4/1/1984	Steam Turbine
7	Station 53	23 604	B	Rochester	7/1/1918	Conventional Hydro

estimated via expected gradients based on path-integrated gradients, which essentially alleviates the computational burden of finding the exactly optimal Shapley values for a deep learning model.

2.5. Error metrics and statistical significance test

To validate model performance, the mean absolute error (MAE) and symmetric mean absolute percentage error (sMAPE) are calculated. The forecast error of model a between target points Y_a and predicted points $\hat{Y}_a$ is defined as: $e_a = Y_a - \hat{Y}_a$. The MAE is defined as:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_a|}{n}.$$

And the sMAPE is defined as:

$$\text{sMAPE} = \frac{100\%}{n} \sum_{a=1}^{n} \frac{|e_a|}{\left(|Y_a| + \left|\hat{Y}_a\right|\right)/2},$$

where n is the length of the temporal sequence.

Because of different random training scenarios and the noisy data properties, the performance of the models could appear different, which may be mistakenly regarded as improvement or superior skill of a model [62]. Therefore, a formal statistical significance test should be conducted between the forecasts. Specifically, the Diebold–Mariano test (DMtest) [63,64] is computed to examine the statistical significance of a model's predictive accuracy. As a model-free method, the DMtest measures the forecasting errors instead of the models themselves. The modified test proposed by Harvey [64] is applied in this work, as it applies to error functions other than quadratic loss. The alternative hypothesis is that the benchmark method is less precise than the comparative model, whereas the null hypothesis is that both approaches produce equally accurate forecasts. The absolute differential performance loss between the proposed model L_p and benchmark model L_b at forecast time t is: $d_t = |L_p - L_b|$. The null and alternative hypotheses are as follows:

$$H_0 : E\left(d_t\right) = 0 \quad \forall t$$

$$H_1 : E\left(d_t\right) > 0 \quad \forall t$$

If a small p-value is observed and lower than the accepted level of 5%, then the null hypothesis H_0 is rejected. In other words, the performance of the proposed model is significantly better than that of the benchmark model.

3. Data and experiment preparation

This section focuses on data retrieval processing, benchmark models introduction, and the experimental procedure for hyper-parameters.

3.1. Data

Eight spatially and technically representative generators were chosen out of 537 historical spot price records, which are available from the NYISO official website [44]. Generator diversity was partially achieved by evenly sampling from the NYISO load subzones. The details of the generators are summarized in Table 1. The proposed forecasting model was trained on a dataset from 2016 to 2018 and validated on data from 2019, to find the neural network structure that was most generalizable across all eight electricity price records.

3.2. Hyper-parameter optimization

An exhaustive grid search was undertaken to acquire the best combination of hyper-parameters that maximize predictive accuracy; their values are shown in Table 2. One key hyper-parameter is the head number in the multi-head attention layers (Head), representing the number of parallel dot-product attention layers. Moreover, the proper dropout rate is also essential, and it should be carefully selected to prevent overfitting. In addition, model variance can also be measured based on a Monte Carlo process [65]. Specifically, the model is repeatedly trained and validated 30 times for each generator, with different random seed settings of the computation system. Then, the final predicted results are assessed based on the mean of the resulting forecasts.

3.3. Benchmark models

Several benchmark models were selected to evaluate the forecasting performance of the proposed model. All the models share the same features to ensure a consistent experimental procedure, except for the SARIMA model, which is a univariate benchmark. The statistical significance of each model's predictive accuracy is compared to the proposed model via the DMtest.

- **SARIMA**: The seasonal auto-regressive integrated moving average model [66] with auto parameters selected based on the minimum Akaike Information Criterion. This model considered only the short-term rolling sliding window and is regarded as the naïve model to measure the performance of the proposed method.
- **Dense**: A naïve neural network model with one fully-connected layer containing 100 units.
- **GRU**: The same GRU model as the proposed ATTnet, with the attention layers removed; it is regarded as a benchmark model to measure the improvement obtained from the multi-head attention layers.
- **HFnet**: A state-of-the-art model with three parallel GRU branches and extra holiday features, proposed by the authors in 2019 [45].
- **GHTnet**: A state-of-the-art parallel CNN-GRU network with the statistical features [46].
- **QCAE**: A state-of-the-art model based on a CNN Autoencoder [67].
- **ATTnet**: The proposed model with multi-head attention layers.

An exhaustive grid search was conducted to obtain the benchmark model hyper-parameters, based on the best general performance across the eight selected NYISO generator datasets. The test range is the same as in Section 3.2. To ensure consistency across the models and experiments, all models (except the SARIMA model, which is regarded as the univariate benchmark) share the same input data variables and sub-sampling rates as the proposed ATTnet.

3.4. Computational environment

The computational environment for model training and validation occurred on an Intel i7 CPU @ 3.80 GHz processor with 16 GB memory and an NVIDIA GeForce GTX 1080 8 GB GPU. The model was developed in Python 3.7 with Tensorflow 2.7.0.

Table 2
ATTnet hyper-parameter optimization,with optimal parameters denoted in bold font.

Layer	Hyper-parameter	Value
Normalization	Type	MinMaxScaler, StandardScaler, **QuantileTransformer**
Multi-Head Attention	Attention Head	1, 2, 3, 4, **5**, 6, 7, 8, 9, 10
Layer Normalization	Epsilon	0.1^3, 0.1^4, 0.1^5, 0.1^6, 0.1^7, 0.1^8, 0.1^9
GRU	Unit	19, **38**, 57, 76
GRU	activation type	**tanh**, relu, elu
Activation	activation type	tanh, **relu**, elu
Dropout 1	Dropout rate	0, 0.01, **0.05**, 0.1, 0.2, 0.3, 0.4, 0.5
Dropout 2	Dropout rate	0, 0.01, **0.05**, 0.1, 0.2, 0.3, 0.4, 0.5
N/A	Batch Size	2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9
N/A	Optimiser	SGD, Adadelta, RMSprop,Adam
N/A	Learning Rate	0.1, 0.01, **0.001**, 0.0001

4. Results and discussion

4.1. General results

Tables 4 and 5 summarize the model performance under three different forecasting horizons. The proposed model, ATTnet, outperforms all the benchmark and state-of-the-art models. As mentioned in Section 2.3, the unobserved windows (GAP) are applied to test the model performance under different energy market bidding and dispatch scenarios. Specifically, there are three types of GAPs for 5-minute, real-time markets: 18, 62 and 144, which represent $18 \times 5 = 90$ minutes, $62 \times 5 = 310$ minutes, and $144 \times 5 = 720$ minutes, respectively. The size of the sliding window sampling is 288, which represents one-day prior; the interval of the long-term moving window is equal to the GAP size for each type of experiment. Data in this time span cannot be observed to inform the predicted price points, due to market restrictions. Model performances are validated via their lower error median line within a narrower interquartile range shown in Fig. 6, which represents the lowest error at relatively low variance.

The results show that adding the multi-head attention mechanism increased the predictive accuracy by 51% in both error metrics compared to the single GRU. In comparison with other cutting-edge models, the proposed model improves on predictive accuracy, as measured by MAE, by an average of 2.80% (HFnet), 2.25% (GHTnet), and 18.00% (QCAE). Similarly for the sMAPE, ATTnet represents an improvement of 4.02% (HFnet), 2.67% (GHTnet), and 16.75% (QCAE) across all prediction gaps. These results demonstrate that the proposed model is superior, as ATTnet is capable of achieving highly accurate forecasts with a less complex model structure. Fig. 7 illustrates forecasts produced by the proposed model and benchmark models.

As Fig. 6 shows, the general performance between ATTnet and two state-of-the-art multiple-branch DL models (HFnet and GHTnet) is relatively close. To ensure the performance improvement is statistically significant, the one-sided DMtest is performed, results of which are reported in Table 3. If a p-value is smaller than 0.05, the null hypothesis will be rejected, which means the average absolute difference between the proposed model and benchmark model is equal to or larger than zero. Most of the computed p-values are near zero (all numbers are rounded to two decimal points to improve readability in the table), which means the proposed model is statistically better than the model to which it is compared. However, several exceptions to this statement occur in the larger forecasting horizon of 144, specifically when ATTnet is compared to GHTnet and HFnet. Generators with the indices of 0, 3, and 4 did not pass the DMtest, which means the performance of the proposed model is equal to or less than the state-of-the-art benchmark model. The main reason for these exceptions will be discussed in the next Section.

4.2. Prediction gap influence

As can be observed in Tables 3, 4, and 5, ATTnet's best performance occurs with a GAP of 18, which means the unobserved prediction horizon is one and a half hours, or 90 min. This is 15 min longer than the required closure of bidding in the NYISO market. Thus, the ATTnet model outperforms the state-of-the-art models in a market environment. As the prediction GAP increases to 62 (5 h), the prediction accuracy reduces on average by 8.8% in MAE and 7.6% in sMAPE among the different models, compared to the smallest GAP. For a GAP size of 144 (12 h), the average reduction in MAE and sMAPE among all the models is 12.8% and 10.0% respectively. While the average performance improves across generators, the proposed model cannot guarantee superior performance at three generators (indices of 0, 3, 4), compared to the state-of-art models of HFnet and GHTnet, at the largest GAP of 144.

Fig. 8 summarizes the attention score when ATTnet is applied to different generator types and different unobserved GAPs are enforced. The x-axis bins the temporal order of the input mini-batch, with the smaller magnitude representing a distance further from the predicted price. The attention score is averaged by generator type, and summed every 48 points with a non-overlapping moving window. Results show that the temporal recurrence is given relatively the same importance along the time axis, for the smaller GAP. This shows that, according to Eq. (1), each point in the input sequence appears with the same likelihood of ‘similarity.’ Conversely, the noted ‘decay importance’ phenomenon [68] can be observed for the larger GAP sizes of 62 and 144. With an increased distance between the points and targets, temporal dependencies are assigned lower attention scores. This is most evident for the largest GAP of 144. The importance score is significantly higher, with a value of 0.5, for the closer temporal index (240 to 288). In addition, an attention score of 0.4 is assigned to the next closest time step, from 192 to 240. This ’decay importance’ means the attention scores assigned to the rest of historical points are divided to sum to 0.1. Thus, their contribution towards the predicted price target is negligible.

Two things can explain this phenomenon. First, due to the large GAP size, the perspective of the auto-regressive process hides the historical observations dependent on the target price in the unobserved window (see Fig. 9), which the model cannot capture. As Tables 4 and 5 show, the ATTnet model performance can be improved by reducing the GAP size. However, a trade-off must be made to leave adequate time for players to submit bids to the market operator. Secondly, because the target points are more dependent on the nearer observations, prices far away from predicted points cannot help update the model to improve weight optimization. The inactivity of GRU following the multi-head attention layers is one of the conclusive proofs: the mean of update gate weights distribution has decreased by over 46%. In other words, with a larger GAP, deep learning is less active in learning the temporal patterns. That is also why the attention score is smaller for the observations farther away from the predicted points.

4.3. Feature importance

As discussed in Section 2.1, the attention score can be interpreted as the probability of similarity in the sequence, giving one measure of feature importance. SHAP is another method of feature importance,

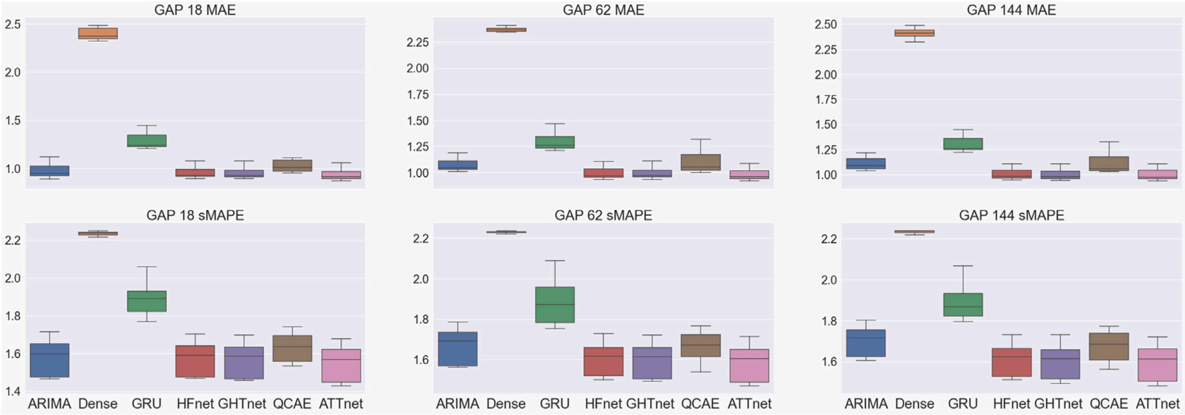

Fig. 6. General results of model performance (log-scale).

Table 3
p-value of DMtest between the proposed ATTnet and benchmark models.

	GAP\Generator	0	1	2	3	4	5	6	7
ATTnet vs. SARIMA	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATTnet vs. Dense	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATTnet vs. GRU	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ATTnet vs. HFnet	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.4	0.00	0.00	0.07	0.11	0.00	0.00	0.00
ATTnet vs. GHTnet	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.21	0.00	0.00	0.01	0.00	0.00	0.00	0.00
ATTnet vs. QCAE	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	144	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4
MAE for the selected generators across forecast horizons (GAP). ATTnet outperforms all benchmark models, for all representative generators, at a GAP of 18, which is 1.5 h ahead. It remains the best performing model at larger forecasting horizons for 75% of generators.

Index	GAP	SARIMA	Dense	GRU	HFnet	GHTnet	QCAE	ATTnet
0	18	9.82	487.72	38.12	9.18	8.93	11.97	**8.62**
1	18	19.85	306.03	28.00	17.96	17.88	19.23	**17.40**
2	18	13.13	277.24	20.38	11.99	11.95	12.96	**11.43**
3	18	9.04	229.31	16.55	8.65	8.62	10.27	**8.48**
4	18	8.19	244.26	16.97	8.11	8.20	9.38	**7.76**
5	18	8.62	209.62	16.89	8.32	8.26	9.33	**7.87**
6	18	8.44	214.62	16.12	8.37	8.32	10.10	**7.93**
7	18	7.79	222.72	17.63	7.85	7.82	8.98	**7.46**
0	62	12.07	457.23	37.84	10.33	**9.79**	15.98	9.84
1	62	23.37	257.21	29.42	18.83	18.70	20.76	**18.35**
2	62	15.56	236.99	20.09	12.76	12.89	14.51	**12.31**
3	62	11.22	224.80	16.29	9.49	9.60	11.63	**9.34**
4	62	10.34	220.62	17.23	8.90	8.94	10.92	**8.68**
5	62	10.84	223.10	17.72	9.03	9.15	10.01	**8.78**
6	62	10.88	234.85	16.91	9.24	9.15	10.61	**8.86**
7	62	10.20	231.20	18.91	8.59	8.62	10.50	**8.36**
0	144	13.66	505.87	38.55	10.46	**10.20**	15.52	10.46
1	144	24.23	308.75	28.00	18.97	19.05	21.30	**18.75**
2	144	16.43	260.31	21.50	12.81	12.73	14.82	**12.72**
3	144	12.28	222.30	17.13	9.72	9.72	11.82	**9.68**
4	144	10.97	256.22	17.93	9.05	9.06	10.96	**9.04**
5	144	11.57	267.12	18.09	9.22	**9.11**	10.67	9.11
6	144	12.29	211.19	16.57	9.48	9.20	10.80	**9.14**
7	144	10.92	248.16	18.32	8.78	8.75	10.89	**8.65**

Table 5
sMAPE for the selected generators.

Index	GAP	SARIMA	Dense	GRU	HFnet	GHTnet	QCAE	ATTnet
0	18	29.82	177.76	114.54	29.91	28.72	36.28	**27.17**
1	18	51.63	173.91	83.41	50.42	49.91	54.96	**47.59**
2	18	44.64	170.39	72.40	42.91	42.43	47.26	**40.89**
3	18	29.91	168.25	58.48	29.34	29.15	35.82	**28.31**
4	18	44.67	174.38	83.81	44.15	44.33	52.36	**42.20**
5	18	36.05	167.86	66.89	35.38	34.84	39.73	**33.45**
6	18	29.20	164.23	64.50	29.41	29.12	34.11	**26.76**
7	18	43.21	174.92	87.29	43.55	42.53	48.74	**41.67**
0	62	37.19	176.89	122.81	33.52	31.42	40.88	**30.75**
1	62	61.00	169.24	93.28	53.60	52.80	58.63	**51.72**
2	62	53.53	169.54	60.13	45.32	45.30	52.33	**44.23**
3	62	36.66	166.12	56.73	32.00	32.13	41.29	**30.93**
4	62	55.07	172.84	80.76	46.90	46.65	55.26	**45.78**
5	62	45.09	170.07	60.93	37.75	37.64	42.26	**36.76**
6	62	36.47	166.96	68.69	31.62	31.21	34.64	**29.50**
7	62	54.11	169.99	90.35	45.38	44.91	52.36	**44.51**
0	144	42.65	180.18	117.12	34.04	32.87	41.17	**32.11**
1	144	63.44	173.07	79.55	54.03	53.97	59.24	**52.70**
2	144	56.45	171.40	67.16	45.72	**45.10**	52.65	45.63
3	144	40.29	165.59	63.78	32.40	32.48	38.65	**31.70**
4	144	56.83	173.80	85.45	47.23	46.93	54.83	**46.50**
5	144	47.82	173.25	68.04	38.69	37.80	44.32	**37.47**
6	144	40.40	166.49	62.49	32.58	31.19	36.62	**30.26**
7	144	56.81	172.89	86.92	45.92	45.00	54.69	**44.97**

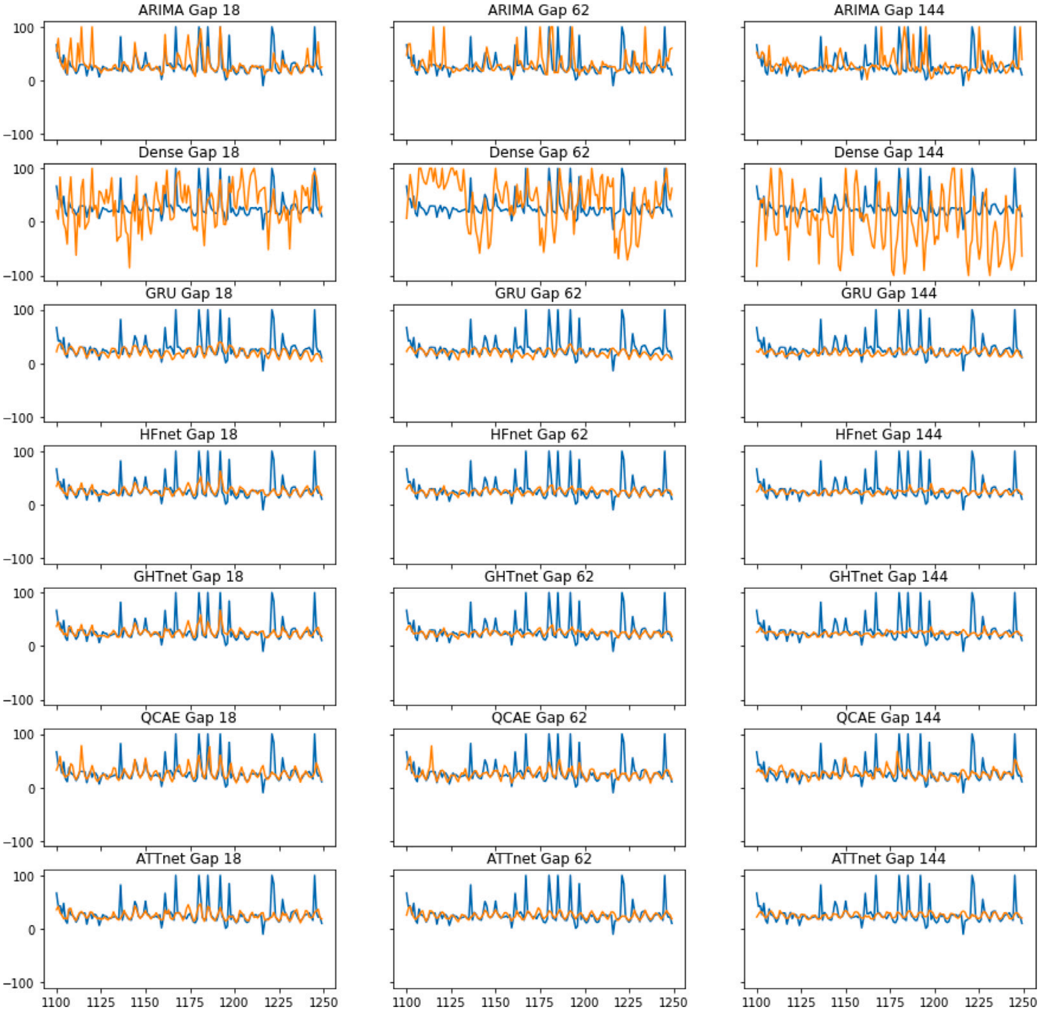

Fig. 7. Bliss Wind Power (Index 1) forecasting examples. Extreme outliers have been removed to improve visualization of the time series. The rate of outlier removal accounts for less than 0.2% of the overall dataset.

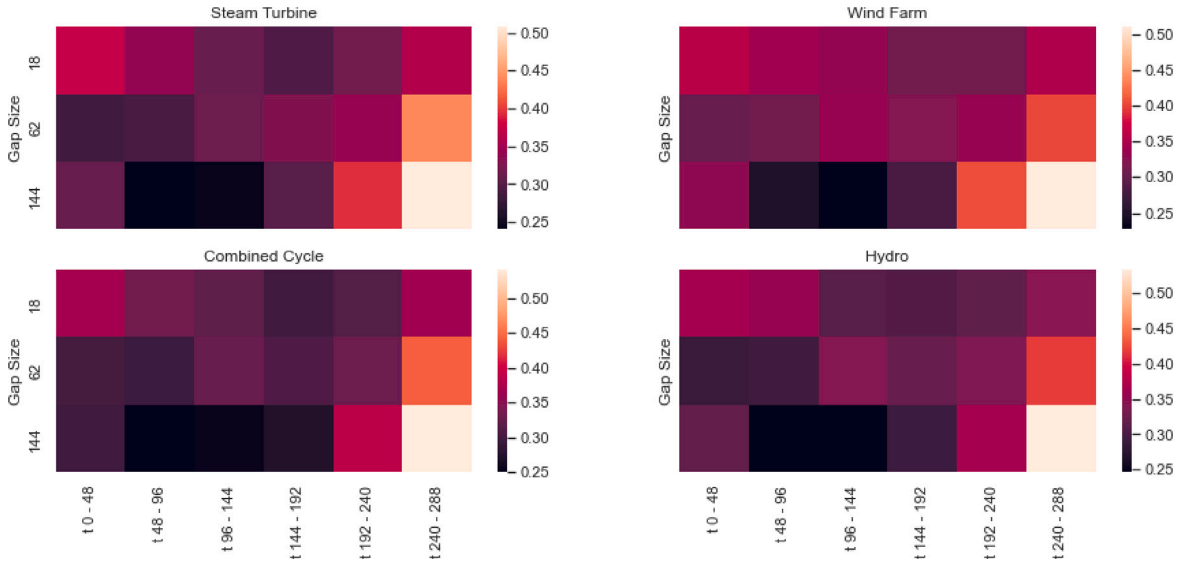

Fig. 8. Attention Score along the temporal axis averaged by generator type. (The attention scores are summed across each 48 interval window.)

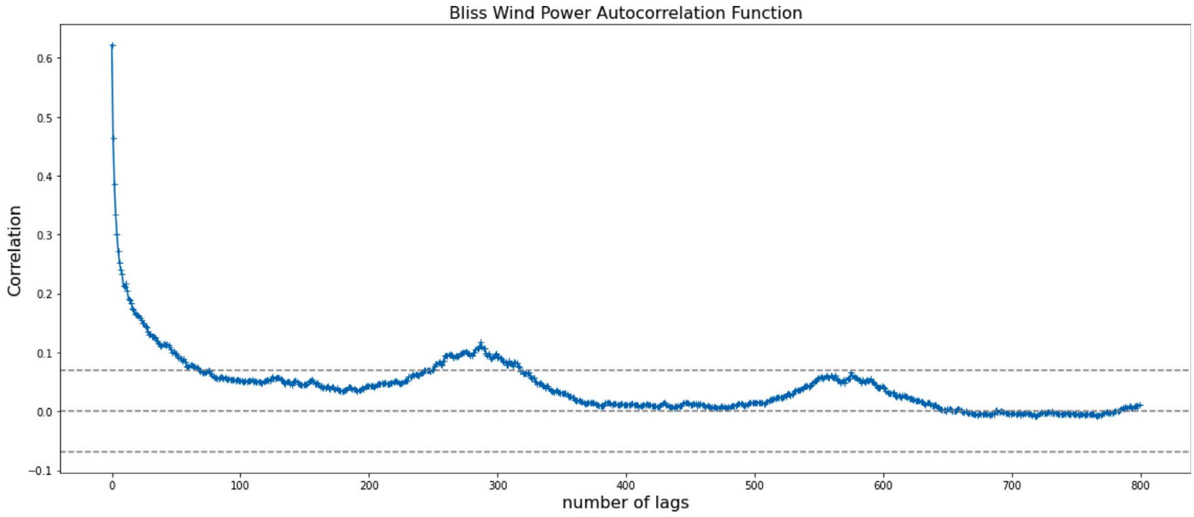

Fig. 9. The ACF plot example.

which can assess the global importance of the feature to all price predictions.

Table 6 shows the mean of the absolute Shapley value for each input feature. The historical LBMP, which is the historical price feature, is the most impactful feature for real-time price prediction, compared to the others. Except for the historical LBMP, the Shapley value distribution of different features is summarized in Fig. 10. In particular, features of temperature, hour, energy bid load, and losses are the next most important features. The box plots for each feature demonstrate the possible range of Shapley values for each feature, according to each target price. For example, the temperature feature sometimes contributes to decreasing the price prediction (Shapley value less than zero), and other times contributes to increasing the predicted price. To improve comprehension of the contribution from different features, a more detailed investigation into each feature needs to be done.

Based on the quantile of the target price, the training set has been classified into ten levels with ascending intervals: Negative Spikes (Q0 and Q1), Negatives Normal (Q2 and Q3), Intermediate Normal (Q4 and Q5), Positive Normal (Q6 and Q7) and Positive Spikes (Q8 and Q9). According to the sampling method, the properties of the input data are divided into 'Long Term' and 'Short Term.' 'Long Term' stands for the input minibatch sampled from the day-interval method, while the 'Short Term' label represents the input data sampled from the sliding rolling window method (see Fig. 3).

As observed in Fig. 11, compared with long-term components, the short-term historical price contributes more to the predicted values. In addition, in the higher or lower quantile range (unlike within the intermediate range) the influence of the historical price appears to have a positive or negative effect on cost. In other words, electricity price spikes are more likely a result of historical extreme positive or negative prices. This phenomenon corroborates the findings in Section 4.2: the dependent components are missing due to an unobserved gap that is enforced in the modeling framework. Therefore, the model must focus only on the temporally nearest sequences, compensating for the missing information in the gap.

Besides LBMP price, of the features related to grid information, Energy Bid Load also needs to be highlighted. The demand-side bids for electricity within the energy market satisfy the customers' energy demand [69]. As one of the important inputs to a Security Constrained Unit Commitment (SCUC) model run by the system operator, Energy Bid Loads are assembled by zone for use in market scheduling [70]. Fig. 12 shows the Shapley values for Energy Bid Load, as well as

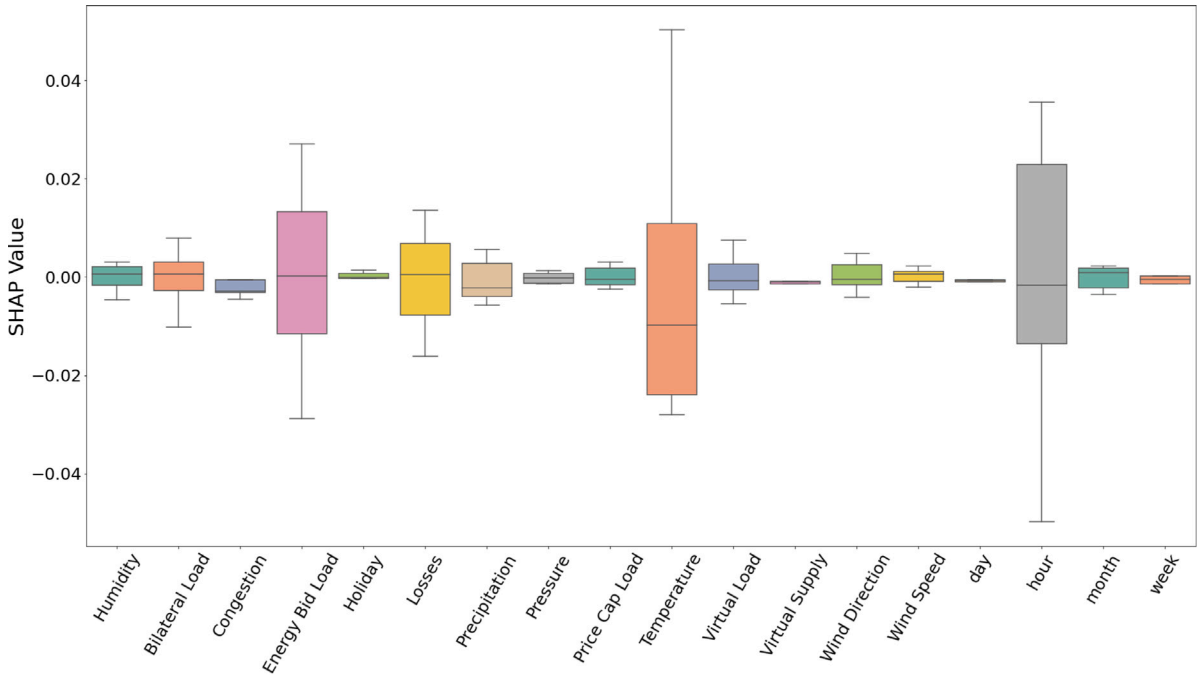

Fig. 10. SHAP: the feature importance.

Table 6
SHAP feature importance. The SHAP value presented below is the mean(abs(SHAP Values)) of each feature across all generators and GAPs.

Class	Feature	SHAP	Unit
Grid Information	**LBMP**	**0.399**	**$/MWh**
	Energy Bid Load	**0.016**	**MW**
	Losses	0.009	$/MWh
	Bilateral Load	0.005	MW
	Congestion	0.005	$/MWh
	Virtual Load	0.004	MW
	Virtual Supply	0.003	MW
	Price Cap Load	0.002	MW
Climate	**Temperature**	**0.025**	**Fahrenheit**
	Precipitation	0.004	Inch
	Wind Direction	0.003	Degree
	Humidity	0.002	%
	Wind Speed	0.001	knots
	Pressure	0.001	Inch
Date	**hour**	**0.025**	**N/A**
	week	0.003	N/A
	month	0.002	N/A
	day	0.001	N/A
	Holiday	0.001	N/A

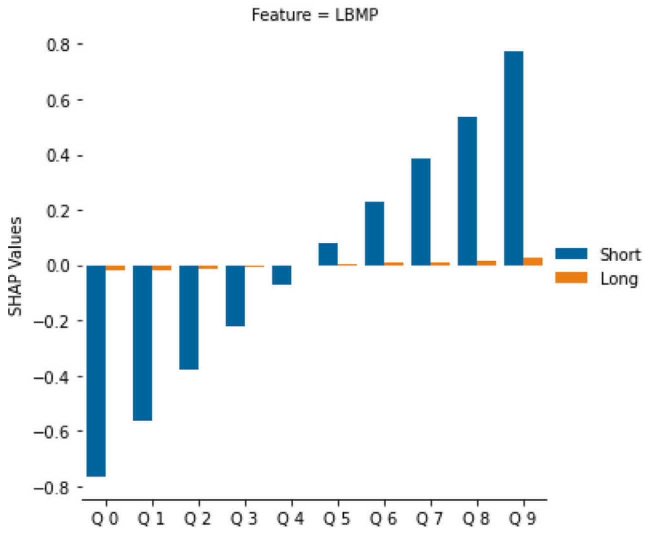

Fig. 11. SHAP: LBMP.

the Losses feature. The short-term Energy Bid Load shares almost the same contribution pattern as the short-term historical price (LBMP). In the short term, this pattern reflects the dynamics of bid load that is aligned with the change in the real-time spot prices. However, its long-term components play the opposite role. This may be caused by an internal interaction between the day-ahead and real-time markets, and more research should be done to shed light on this effect. The Losses (marginal cost losses [$/MWh]) are a feature that reflects the actual market's power price with dynamic load flow [69]. The Losses follow the same trend in long-term and short-term dependencies. Because the marginal cost losses are one of the components of figuring out the final LBMP, the Shapley values for both LBMP, and Losses share the same pattern across multiple quantiles.

ATTnet mainly captures the long-term patterns of the hour and temperature features, along the temporal axis. The long-term effect of Temperature (see Fig. 13) is a higher Shapley value in positive spike ranges (i.e. Q7–Q9). In addition, compared to the negative range, the long-term temperature feature impacts the positive predicted price more. Similar to the temperature feature, the long-term components of the hour (See Fig. 14) contribute much more to the final predicted prices than the short-term, especially within the negative range (Q0–Q3). In other words, the energy market participants' activities towards the dynamics of hour and temperature could delay the impact on electricity price change.

5. Conclusions and future work

A new deep learning model has been developed using multi-head attention to highlight important input features to a GRU neural network architecture. The model, ATTnet, has been validated on eight generators from the NYISO spot price market, demonstrating that it achieves more accurate electricity price predictions than benchmark

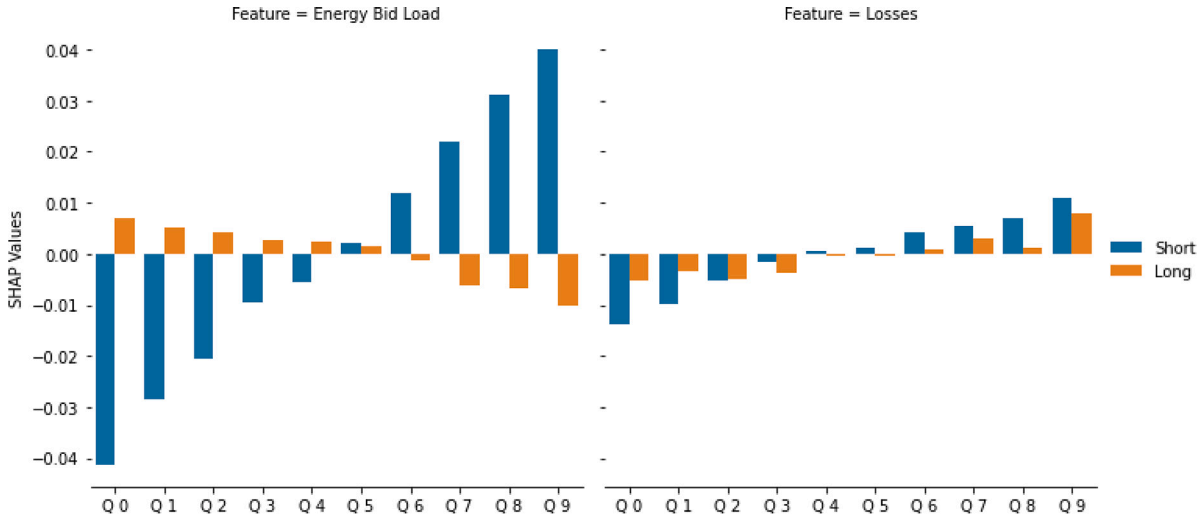

Fig. 12. SHAP: Grid information.

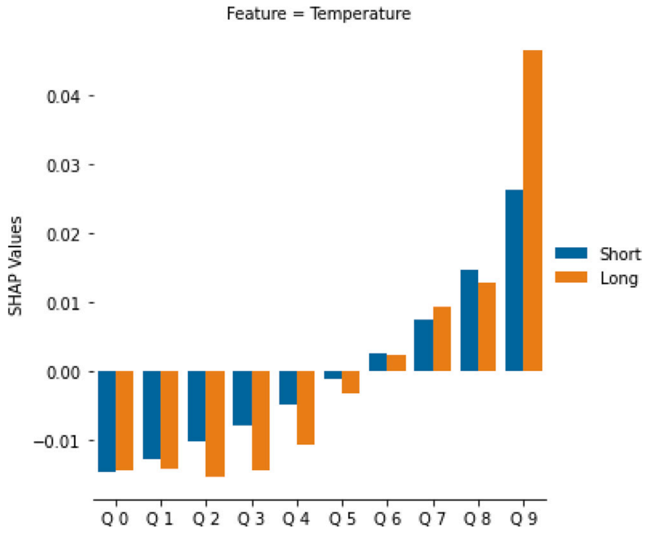

Fig. 13. SHAP: Temperature.

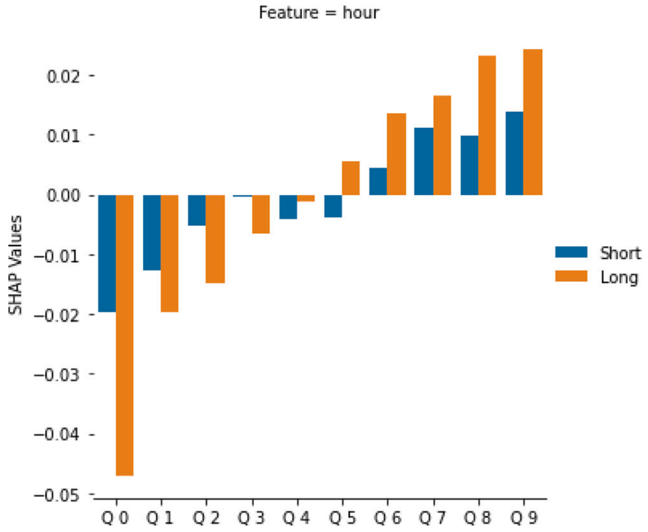

Fig. 14. SHAP: Hour.

and state-of-the-art methods. The novel architecture has two main components - a Residual Multi-head Attention (RMA), followed by a Recurrent Post-Activation Network (RPN). In the RMA, the likelihood of the temporal similarity is captured by the attention mechanism, and the temporal dependencies are input and further processed by the RPN. The proposed model, ATTnet, significantly outperforms other state-of-the-art models (HFnet, GHTnet, and QCAE) with an average 7.69% decrease in MAE and a 7.81% reduction in sMAPE. The attention score is used to evaluate the impact of the information gap size on the model's performance. At larger gap sizes (e.g. 144), the phenomenon of "decay importance" can be observed. That is, the low attention scores attributed to historical data further back in time show that these inputs are less important to the predicted price target. Feature importance is also examined using the SHAP approach. The historical pricing feature is the most significant element for real-time price forecasting, consistent with attention score results. In addition, the effects of grid information (Energy Bid Load and Losses) and other factors (Temperature and Hour) are quantified.

As future work, the authors are actively establishing deep learning networks based on a dilated CNN, to replace the GRU layers, which will save computational resources. In addition, post-calibration modules are also under development to help improve the forecasting performance for the upper and lower quantiles, i.e. the price spike ranges. Finally, a more sophisticated tuning method, such as Bayesian tuning via a Gaussian process, will be applied to facilitate finding the best global combination of hyper-parameters, instead of the current method of tuning these parameters manually via grid search or random search.

CRediT authorship contribution statement

Haolin Yang: Methodology, Investigation, Software, Validation, Visualization, Writing – original draft. **Kristen R. Schell:** Supervision, Conceptualization, Writing – review & editing, Project administration, Resources.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] Conejo AJ, Plazas MA, Espinola R, Molina AB. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. IEEE Trans Power Syst 2005;20(2):1035–42.

[2] Jakaša T, Andročec I, Sprčić P. Electricity price forecasting—ARIMA model approach. In: 2011 8th international conference on the European energy market. EEM, IEEE; 2011, p. 222–5.

[3] Uniejewski B, Marcjasz G, Weron R. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. Int J Forecast 2019;35(4):1533–47.

[4] Weron R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. 2014, http://dx.doi.org/10.1016/j.ijforecast.2014.08.008.

[5] Wu L, Shahidehpour M. A hybrid model for day-ahead price forecasting. IEEE Trans Power Syst 2010;25(3):1519–30.

[6] Smolen J, Dudic B. The role of residual demand in electricity price analysis and forecasting: case of Czech electricity market. Int J Energy Econ Policy 2017;7(5):152.

[7] Sgarlato R, Ziel F. The role of weather predictions in electricity price forecasting beyond the day-ahead horizon. IEEE Trans Power Syst 2022.

[8] Meng A, Wang P, Zhai G, Zeng C, Chen S, Yang X, et al. Electricity price forecasting with high penetration of renewable energy using attention-based LSTM network trained by crisscross optimization. Energy 2022;254:124212.

[9] Feijoo F, Silva W, Das TK. A computationally efficient electricity price forecasting model for real time energy markets. Energy Convers Manage 2016;113:27–35.

[10] Yang H, Schell KR. QCAE: A quadruple branch CNN autoencoder for real-time electricity price forecasting. Int J Electr Power Energy Syst 2022;141:108092.

[11] Tschora L, Pierre E, Plantevit M, Robardet C. Electricity price forecasting on the day-ahead market using machine learning. Appl Energy 2022;313:118752.

[12] Pfenninger S, Hirth L, Schlecht I, Schmid E, Wiese F, Brown T, et al. Opening the black box of energy modelling: Strategies and lessons learned. Energy Strategy Rev 2018;19:63–71.

[13] Dong W, Chen X, Yang Q. Data-driven scenario generation of renewable energy production based on controllable generative adversarial networks with interpretability. Appl Energy 2022;308:118387.

[14] Zhang S, Zhang D, Qiao J, Wang X, Zhang Z. Preventive control for power system transient security based on xgboost and DCOPF with consideration of model interpretability. CSEE J Power Energy Syst 2020;7(2):279–94.

[15] Toubeau J-F, Bottieau J, Wang Y, Vallee F. Interpretable probabilistic forecasting of imbalances in renewable-dominated electricity systems. IEEE Trans Sustain Energy 2021.

[16] Santos OL, Dotta D, Wang M, Chow JH, Decker IC. Performance analysis of a DNN classifier for power system events using an interpretability method. Int J Electr Power Energy Syst 2022;136:107594.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017, CoRR arXiv:1706.03762, URL http://arxiv.org/abs/1706.03762.

[18] Liang Y, Ke S, Zhang J, Yi X, Zheng Y. Geoman: Multi-level attention networks for geo-sensory time series prediction. In: IJCAI. 2018, 2018, p. 3428–34.

[19] Ding Y, Zhu Y, Feng J, Zhang P, Cheng Z. Interpretable spatio-temporal attention LSTM model for flood forecasting. Neurocomputing 2020;403:348–59.

[20] Zhao H, Wang Y, Duan J, Huang C, Cao D, Tong Y, et al. Multivariate time-series anomaly detection via graph attention network. In: 2020 IEEE international conference on data mining. ICDM, IEEE; 2020, p. 841–50.

[21] Fan C, Zhang Y, Pan Y, Li X, Zhang C, Yuan R, et al. Multi-horizon time series forecasting with temporal attention learning. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2527–35.

[22] Gangopadhyay T, Tan SY, Jiang Z, Meng R, Sarkar S. Spatiotemporal attention for multivariate time series prediction and interpretation. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing. ICASSP, IEEE; 2021, p. 3560–4.

[23] Guo S, Lin Y, Feng N, Song C, Wan H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 33, (01):2019, p. 922–9.

[24] Du S, Li T, Yang Y, Horng S-J. Multivariate time series forecasting via attention-based encoder–decoder framework. Neurocomputing 2020;388:269–79.

[25] Hollis T, Viscardi A, Yi SE. A comparison of LSTMs and attention mechanisms for forecasting financial time series. 2018, arXiv preprint arXiv:1812.07699.

[26] Wang S, Wang X, Wang S, Wang D. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. Int J Electr Power Energy Syst 2019;109:470–9.

[27] Fahim SR, Sarker Y, Sarker SK, Sheikh MRI, Das SK. Self attention convolutional neural network with time series imaging based feature extraction for transmission line fault detection and classification. Electr Power Syst Res 2020;187:106437.

[28] Heidari A, Khovalyg D. Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition. Sol Energy 2020;207:626–39.

[29] Niu Z, Yu Z, Tang W, Wu Q, Reformat M. Wind power forecasting using attention-based gated recurrent unit network. Energy 2020;196:117081.

[30] Zhang S, Chen Y, Xiao J, Zhang W, Feng R. Hybrid wind speed forecasting model based on multivariate data secondary decomposition approach and deep learning algorithm with attention mechanism. Renew Energy 2021;174:688–704.

[31] Li Y, Tong Z, Tong S, Westerdahl D. A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation. Sustainable Cities Soc 2022;76:103481.

[32] Lin J, Ma J, Zhu J, Cui Y. Short-term load forecasting based on LSTM networks considering attention mechanism. Int J Electr Power Energy Syst 2022;137:107818.

[33] Qin J, Zhang Y, Fan S, Hu X, Huang Y, Lu Z, et al. Multi-task short-term reactive and active load forecasting method based on attention-LSTM model. Int J Electr Power Energy Syst 2022;135:107517.

[34] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems. vol. 30, Curran Associates, Inc.; 2017, URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[35] Zhang K, Zhang J, Xu P-D, Gao T, Gao DW. Explainable AI in deep reinforcement learning models for power system emergency control. IEEE Trans Comput Soc Syst 2021.

[36] Zarbakhsh N, Misaghian MS, McArdle G. Human mobility-based features to analyse the impact of COVID-19 on power system operation of Ireland. IEEE Open Access J Power Energy 2022.

[37] Hoffmann V, Klemets JRA, Torsæter BN, Rosenlund GH, Andresen CA. The value of multiple data sources in machine learning models for power system event prediction. In: 2021 international conference on smart energy systems and technologies. SEST, 2021, p. 1–6. http://dx.doi.org/10.1109/SEST50973.2021.9543226.

[38] Liu J, Zhang Z, Fan X, Zhang Y, Wang J, Zhou K, et al. Power system load forecasting using mobility optimization and multi-task learning in COVID-19. Appl Energy 2022;310:118303.

[39] Khan W, Walker S, Zeiler W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. Energy 2022;240:122812.

[40] Park JH, Jo HS, Lee SH, Oh SW, Na MG. A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. Nuclear Eng Technol 2022;54(4):1271–87.

[41] Li W, Becker DM. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. Energy 2021;237:121543.

[42] Graves A, Wayne G, Danihelka I. Neural turing machines. 2014, arXiv preprint arXiv:1410.5401.

[43] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014, arXiv preprint arXiv:1409.0473.

[44] NYISO. 2019 State of the market report for the New York ISO markets. 2020, https://www.nyiso.com/documents/20142/2223763/NYISO-2019-SOM-Report-Full-Report-5-19-2020-final.pdf/bbe0a779-a2a8-4bf6-37bc-6a748b2d148e?t=1589915508638.

[45] Yang H, Schell KR. HFNet: Forecasting real-time electricity price via novel GRU architectures. In: 2020 international conference on probabilistic methods applied to power systems. PMAPS, IEEE; 2020, p. 1–6.

[46] Yang H, Schell KR. GHTnet: Tri-branch deep learning network for real-time electricity price forecasting. Energy 2022;238:122052.

[47] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[48] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.

[49] Ba JL, Kiros JR, Hinton GE. Layer normalization. 2016, arXiv preprint arXiv:1607.06450.

[50] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014, arXiv preprint arXiv:1412.3555.

[51] Staudemeyer RC, Morris ER. Understanding LSTM – a tutorial into long short-term memory recurrent neural networks. 2019, arXiv:1909.09586.

[52] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.

[53] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. 2021, arXiv:2012.07436.

[54] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems 30. Curran Associates, Inc.; 2017, p. 4765–74, URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[55] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):2522–5839.

[56] Mitchell R, Frank E, Holmes G. GpuTreeShap: massively parallel exact calculation of SHAP scores for tree ensembles. PeerJ Comput Sci 2022;8:e880.

[57] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2(10):749.

[58] Hart O, Moore J. Property rights and the nature of the firm. J Polit Econ 1990;98(6):1119–58.

[59] García MV, Aznarte JL. Shapley additive explanations for NO2 forecasting. Ecol Inform 2020;56:101039.

[60] Lundberg S. Shap.GradientExplainer. 2018, https://shap-lrjball.readthedocs.io/en/latest/generated/shap.GradientExplainer.html.

[61] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Precup D, Teh YW, editors. Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol. 70, PMLR; 2017, p. 3319–28, URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[62] Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Appl Energy 2018. http://dx.doi.org/10.1016/j.apenergy.2018.02.069.

[63] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat 2002;20(1):134–44.

[64] Harvey D, Leybourne S, Newbold P. Testing the equality of prediction mean squared errors. Int J Forecast 1997;13(2):281–91.

[65] Zhu L, Laptev N. Deep and confident prediction for time series at uber. In: 2017 IEEE international conference on data mining workshops. ICDMW, IEEE; 2017, p. 103–10.

[66] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2018.

[67] Yang H, Schell KR. QCAE: A quadruple branch CNN autoencoder for real-time electricity price forecasting. Int J Electr Power Energy Syst 2022;141:108092. http://dx.doi.org/10.1016/j.ijepes.2022.108092, URL https://www.sciencedirect.com/science/article/pii/S014206152200134X.

[68] Ziel F. Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. IEEE Trans Power Syst 2016;31(6):4977–87. http://dx.doi.org/10.1109/TPWRS.2016.2521545.

[69] NYISO. Market participants user's guide. 2021, https://www.nyiso.com/documents/20142/3625950/mpug.pdf.

[70] NYISO. Day-ahead scheduling manual. 2021, https://www.nyiso.com/documents/20142/2923301/dayahd_schd_mnl.pdf/0024bc71-4dd9-fa80-a816-f9f3e26ea53a.