Forecasting Nodal Price Difference Between Day-Ahead and Real-Time Electricity Markets Using Long-Short Term Memory and Sequence-to-Sequence Networks

RONIT DAS, (Student Member, IEEE), RUI BO, (Senior Member, IEEE), HAOTIAN CHEN, (Student Member, IEEE), WAQAS UR REHMAN, (Graduate Student Member, IEEE), AND DONALD WUNSCH II, (Fellow, IEEE)
Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA
Corresponding author: Rui Bo (rbo@mst.edu)

ABSTRACT Price forecasting is at the center of decision making in electricity markets. Much research has been done in forecasting energy prices for a single market while little research has been reported on forecasting price difference between markets, which presents higher volatility and yet plays a critical role in applications such as virtual trading. To this end, this paper takes the first attempt at it and employs novel deep learning architecture with Bidirectional Long-Short Term Memory (LSTM) units and Sequence-to-Sequence (Seq2Seq) architecture to forecast nodal price difference between day-ahead and real-time markets. In addition to value prediction, these deep learning architectures are also used to develop classification models to predict the price difference bands/ranges. The proposed methods are tested using historical PJM market data, and evaluated using Root Mean Squared Error (RMSE) and other customized performance metrics. Case studies show that both deep learning methods outperform common methods including ARIMA, XGBoost and Support Vector Regression (SVR) methods. More importantly, the deep learning methods can capture the magnitude and timing of price difference spikes. Numerical results show the Seq2Seq model performs particularly well and demonstrates generalization capability to extended forecasting lead time.

INDEX TERMS Electricity markets, DA/RT price difference, forecasting, long-short term memory, LSTM, sequence to sequence, Seq2Seq, deep learning.

I. INTRODUCTION

With the development of competitive deregulated electricity markets, electricity price forecasting has become crucial for market participants: precise short-term forecast of electricity prices is decisive for generation companies (GENCOs) to develop optimal bidding strategies and participate in electricity markets, and thereby maximize their revenue. Medium-term forecast of electricity prices is exploited by power

The associate editor coordinating the review of this manuscript and approving it for publication was Elizete Maria Lourenco.

producers to commit to favorable bilateral contracts. In the long-term, GENCOs leverage the electricity price forecasting to make investment decisions for maintenance planning and generation capacity planning. Moreover, short-term and medium-term forecast of electricity prices is utilized by large consumers and load serving entities (LSEs) to optimally bid in the electricity market and enter into favorable bilateral contracts [1]–[11].

Market clearing price (MCP) is the fundamental pricing concept in wholesale electricity markets, and there is only one MCP for the entire power network in the absence of

transmission congestion; but for a congested network, the electricity price may be dissimilar at different locations, and thus, termed as locational marginal price (LMP) [6]. Volatility is the most distinct characteristic of the electricity price: dissimilar to load, the electricity price is non-homogeneous and displays minimal cyclic property which renders it challenging to forecast [2]. The price volatility and spikes can be instigated by many factors, including load uncertainties, generation uncertainties, transmission network congestion, fluctuating fuel prices, market participants' behavior, and weather conditions [2].

LMPs in day-ahead (DA) market and real-time (RT) market can vary substantially due to price volatility itself as well as the discrepancy between forecast market conditions in the DA market and the realized uncertainties in the RT market. Therefore, compared to the volatility of prices in one market, the price differences between DA and RT markets, are extremely volatile and difficult to forecast.

A. MOTIVATION

DA/RT price difference can be valuable for various applications. The most common examples are Financial Transmission Right (FTR) auctions market and virtual bidding. The FTR auctions - a financial derivative market - aim at hedging against transmission congestion cost, and therefore offer price certainty to market participants [12]. Virtual bidding was introduced in the electricity markets to enhance the DA/RT price convergence [13]. Virtual bidders (VBs) can participate as supply (or demand), based on LMP predictions, in DA market and correspondingly settle as demand (or supply) in the RT market with the same amount of energy. VBs' revenues largely hinge upon the DA/RT price difference [14]. To the best of authors' knowledge, prediction of nodal market price difference between DA and RT markets has not been reported in the extant literature. Hence, novel forecasting methods using Bidirectional Long-Short Term Memory networks (LSTM) and Sequence-to-Sequence networks are proposed in this work to predict DA/RT nodal LMP difference, or cross-market nodal price difference. The Bidirectional LSTMs and Seq-2-Seq models are specialized variants of the regular LSTM [15]. LSTMs have the ability to model long-term, as well as, short-term complex dependencies in sequences thereby making them a suitable approach to solve the problem at hand. It is envisioned that VBs can exploit the forecasted DA/RT price difference to strategically bid in electricity markets and increase their likelihoods of making profits in virtual trading, which will aid both markets to achieve better convergence.

B. RELATED WORK

In the existing literature, prediction of electricity prices has been attempted using classical statistical methods like Autoregressive Integrated Moving Average (ARIMA) models [16], Wavelet Transform [17], etc.

With recent advancements in AI, especially on neural networks, Artificial Neural Nets (ANNs) and ensemble methods have also been leveraged for electricity price forecasting [2], [7]–[11]. One of the earlier approaches taken to predict time series values has been demonstrated by [18], where the authors have used a bi-directional approach to simple recurrent neural networks for forecasting missing time series values. This approach is a precursor to the Bidirectional LSTM method which is currently used for various prediction purposes.

LMP Forecasting is a well-researched topic in the extant literature, however forecasting price difference across markets or price spread within a market has rarely been investigated. Lately, a modified version of GARCH model to forecast the volatility of price spread values in Day-ahead market has been reported in [19]. To date, very little research has been conducted on forecasting the DA/RT price difference. Built upon the foundational work in [20], this paper comprehensively reports the pioneering work on DA/RT price difference forecast using proposed state-of-the-art models.

C. CHALLENGES IN FORECASTING DA/RT PRICE DIFFERENCE

The nodal price difference between the DA and RT markets is much more volatile than DA market price and RT market price. Unlike DA LMP or RT LMP, the DA/RT price difference frequently fluctuates between positive values and negative values. It can be seen from Fig.1 which shows the nodal DA/RT price difference at 02BRUSH bus of PJM market in 2018. In this figure, there are sporadic high spikes occurring over the course of the year with no clear patterns observed. It should also be noted many of the spikes occurred for just one or very few hours, and therefore presents particular challenge in forecasting the timing of those spikes. However, the occurrence of high spikes is of great importance as it indicates larger deviation between DA and RT markets, more financial opportunity for market participants, and higher need for better convergence of the two markets. Having a good forecasting capability for the DA/RT price difference will help promote liquidity of the markets and achieve better market convergence. While cyclic patterns may exist in both DA market price and RT market price, it is much less present in DA/RT price difference, making the DA/RT price forecast a more challenging problem, which can be observed from auto-correlation analysis results on the market data. Fig 1b, Fig.1c and Fig.1d respectively show the auto-correlation values for DA/RT price differences, DA prices, and RT prices of 168 hours (i.e., one week). Therefore, forecasting the price difference values is very challenging, even for single period forecasting (e.g., the next hour).

Moreover, such highly volatile time series many times contains hard to find non-linear patterns among different time steps. We conduct the Terasvirta Neural Network Test to determine the statistical significance of the presence of non-linear patterns in our time series. The hypotheses for the test is as follows:
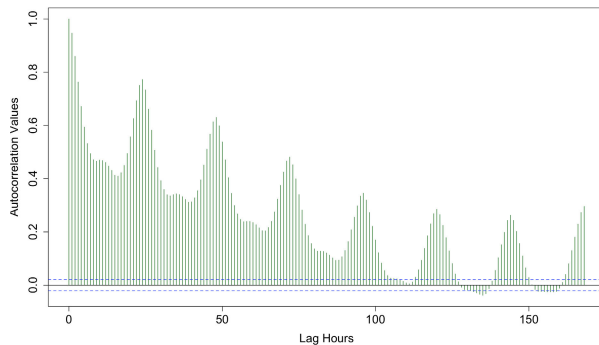
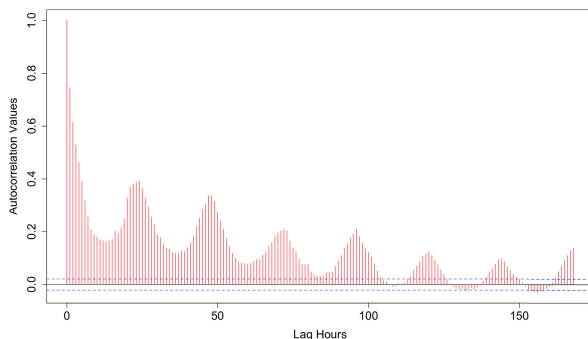H_0: Linearity

H_a: Non-linearity
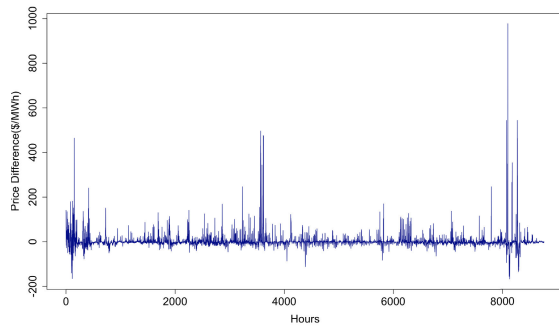
(a) Historical DA/RT price difference at node 02BRUSH of PJM market in 2018
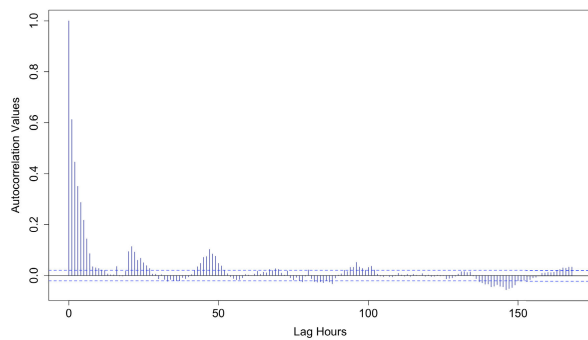
(b) Auto-correlation of DA/RT price differences

(c) Auto-correlation of DA prices

(d) Auto-correlation of RT prices

FIGURE 1. Auto-correlation analysis on market prices.

The Terasvirta Neural Network Test is used to detect neglected non-linearity in time series, using a single-layer feed-forward network and using the logistic function as the

TABLE 1. Terasvirta NN test for non-linearity.

Test Statistic	P-Value
Chi-squared	$< 2.2e^{-16}$
F	$< 2.2e^{-16}$

activation function [21]. The results of the test is shown in the following table1:

As we observe in Table1, we reject the null hypothesis at 95% confidence. Therefore, we conclude by saying that our observed price difference time series has non-linear patterns in it.

D. CONTRIBUTIONS AND PAPER STRUCTURE

The contribution of this work includes:

1) To the best of authors' knowledge, this work is the first in studying DA/RT price difference forecast, which is a more challenging problem than price forecast while having immense value in electricity market applications.

2) Owing to the extremely volatile nature of the day-ahead and real-time price difference values and also our goal of predicting the spikes in the values, deep learning methods, especially Bidirectional Long-Short Term Memory and Sequence-to-Sequence models are developed to tackle the forecasting problem since these models are able to capture complex patterns in the data in order to make accurate forecast [22]. In particular, this work is the first in applying Sequence-to-Sequence network architecture to market price forecast. In addition to value forecast, the deep learning architectures are also applied to perform band forecast.

3) Comprehensive performance evaluation is proposed that not only evaluates the performance in forecasting the values, but also in the direction of price difference, spike value, and particularly important, the timing of the spike values. It should be stressed that this work imposes a stricter performance measure to evaluate whether the forecast can capture the sign of price difference values, and the timing of price spikes, which has not been reported in the literature and yet is particularly valuable for applications such as virtual bidding.

4) Through numerical comparison with common forecast methods such as ARIMA, XGBoost, SVR, the proposed deep learning models demonstrated its superior performance under all evaluation metrics. In particular, its capability in forecasting the magnitude of spike values and their timing is very encouraging.

5) The proposed deep learning models are also adjusted to do band forecast, which is a classification problem. Superior performance is also observed, demonstrating the ability of the proposed architecture in both value forecast and band forecast.

The rest of the paper is organized as follows. Section II introduces deep learning architectures, Bidirectional LSTM and Seq2Seq, and introduces the proposed architecture for the DA/RT price difference forecast problem as well as the band forecast problem. Section III briefly reviews a few common forecasting methods, and then extensively discusses the

performance evaluation methods. A comprehensive case study is presented in Section IV, where the performance of common methods and proposed methods is examined by various evaluation metrics in great detail. Section V concludes the paper.

II. METHODOLOGY

This section gives a brief introduction of the different types of Recurrent Neural Nets(RNNs), specifically, Bidirectional LSTMs and Seq2Seq network architectures, used in time series forecasting and also explains why these neural network architectures work in order to make better forecasts.

A. MOTIVATION FOR USING BI-DIRECTIONAL LSTM AND SEQUENCE-2-SEQUENCE NETWORKS FOR TIME SERIES FORECASTING

In order to understand how Long-Short Term Memory networks are able to perform such accurate forecasts for time series problems, one needs to understand how LSTMs work in general. LSTMs are a special type of recurrent neural networks, developed by [15]. Like regular Recurrent Neural Networks, LSTM networks also work on a sequential data and find both short-term and long-term dependencies required to forecast. As RNNs suffer from the vanishing-gradient problem [23] while working on long-term dependencies, LSTM networks mitigate this issue. LSTM networks have internal memory state to remember past patterns which are then used to forecast the future values. While models like ARIMA also take into account the past trends, these are linear models having implied Gaussian assumptions and are only able to model linear dependencies in a time series [24]. Moreover, tree-based models like XGBoost are also used for forecasting in time series problems, but LSTMs prove to be a better alternative owing to their ability to extract complex non-linear features across different time-steps which both the linear models and tree-based models are not able to perform well.

Fig.1a, shows the highly volatile nature of the time series and also the existence of strong non-linear patterns, as discussed in subsection C under Section I. Linear time series models like ARIMA only take into account dependencies which are a few time steps into the past in order to model complex time series; more complex neural network models like LSTMs can better model both long-term and short-term non-linear dependencies, resulting in more accurate predictions [22].

Bi-directional LSTM models, proposed in this paper, build on regular LSTM networks in a way that it consists of 2 recurrent blocks which are trained simultaneously. One block works on the forward sequence, while the other on the same sequence, but reversed in a chronological order. The main intuition behind such an approach is that the network is presented with both the past sequence of values and future sequence, as well, allowing it to learn complex dependencies from the reversed-sequence which a regular LSTM network would not be able to perform. Especially for complex and volatile time series like the one presented in this paper,

having a network being capable to extract complex non-linear features not only from the past sequence, but from the future as well, gives it an additional advantage over other models.

In the same way, Sequence-2-Sequence architectures are encoder-decoder networks which also build on regular RNN/LSTM networks. However, in these models, the encoder model consists of 1 or more layers of LSTM neurons which encode the incoming sequence into an Encoded State. In natural language processing parlance, it is also known as the Context Vector. The Encoded State encapsulates all the important and relevant input features with which the decoder unit forecasts the future values with high degree of accuracy.

Both Bi-directional LSTMs and Sequence-2-Sequence networks offer state-of-the-art unique approaches to model complex and volatile time series problems in order to make accurate forecasts.

Additionally, such complex structures like Bidirectional LSTMs and Sequence-2-Sequence networks also have the potential to extract complex features across different timesteps, not only to make accurate forecast of values or probabilities of different price bands, but of spikes (abrupt increase or decrease in price difference value) in the time series, which makes the application of such state-of-the-art networks to this complex problem a unique and novel contribution to the field of electricity market forecast.

B. ARCHITECTURES FOR BIDIRECTIONAL LSTM AND SEQUENCE-2-SEQUENCE MODELS

Bidirectional and Sequence-2-Sequence models work on an input sequence of certain length. In time series analysis, it is also called the lag, or the amount of time-steps to consider from the past, in order to make predictions in the future. From Fig.1b, we observe that the cut-off for the DA/RT price difference auto-correlation plot is around 48 hours. Hence, for investigating different architectures, we use a lag size of 48 hours.

For neural networks, coming up with an architecture which gives the desired results is a difficult task since it requires a lot of time to train the different models and also fine tune them in order to get the best results. In this paper, different architectures of bidirectional LSTM and Sequence-2-Sequence networks for both value forecast and band forecast, have been explored and the following subsections discuss about them in details.

C. BIDIRECTIONAL LSTM NETWORK ARCHITECTURE

A Bidirectional LSTM network introduced in [20] is implemented in this work. Different combination of hidden layers with LSTM neurons are explored, and the ideal model is selected based on the RMSE value of that model. The proposed architecture and design parameters of the model are described in the Tables 2, 3, 4.

As we observe from the above tables, increasing the number of hidden layers, results in an improvement of the test set RMSE value. It follows from this idea that increasing

TABLE 2. Bidirectional LSTM with 1 hidden layer.

Hidden Layers	Neurons(forward & reversed sequence)
Hidden Layer 1	50 + 50
Output Layer	1
Loss Function	L2(Least Square Errors)
Gradient Descent Optimizer	Adam
Batch Size	64
Test RMSE	41.358

TABLE 3. Bidirectional LSTM with 2 hidden layers.

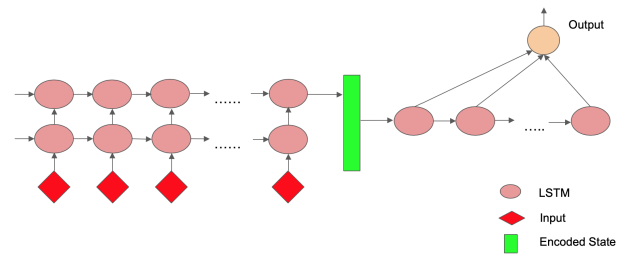| Hidden Layers	Neurons(forward & reversed sequence)
Hidden Layer 1	128 + 128
Hidden Layer 2	64 + 64
Output Layer	1
Loss Function	L2(Least Square Errors)
Gradient Descent Optimizer	Adam
Batch Size	64
Test RMSE	35.19

TABLE 4. Bidirectional LSTM with 3 hidden layers.

Hidden Layers	Neurons(forward & reversed sequence)
Hidden Layer 1	128 + 128
Hidden Layer 2	64 + 64
Hidden Layer 3	10 + 10
Output Layer	1
Loss Function	L2(Least Square Errors)
Gradient Descent Optimizer	Adam
Batch Size	32
Holdout RMSE	33.15

the number of hidden layers acts as a filter to filter out the noise in the data, thereby giving more accurate predictions. The lag, or the lookback window is determined from Fig.1b and hence is kept fixed. Moreover, increasing the number of hidden layers to a more dense structure, also results in a more complex model with a higher probability of overfitting to the training data. And increasing the number of layers increases the number of weights thereby also increasing the training time.

Decreasing the batch size for the training data set also seems to have a positive effect in increasing the accuracy of the prediction by reducing the RMSE in the test set. Increasing the batch size can sometimes speed up the training process, but also tend to get stuck at a sub-optimal local minimum. Therefore having a smaller batch size has the advantage of adding some noise-like characteristics to the training data whereby the algorithm is able to escape such situations and converge at a better optimum point on the error landscape.

The optimizer algorithm, Adam, is used for all the different architectures since it has proven itself to be a reliable algorithm comprising of the benefits from RMSprop and Nesterov-Accelerated Gradient algorithms.

As a result of overfitting, a model captures the noises present in the training set during training, which results in very poor generalizability. In order to prevent that, we apply regularization technique [25].

FIGURE 2. A representational diagram of a Seq2Seq model.

TABLE 5. Seq-2-Seq architectures with 1 hidden layer in the encoder.

Hidden Layers	Neurons
Hidden Layer 1(Encoder)	200
Hidden Layer 1(Decoder)	20
Output Layer	1
Loss Function	L2(Least Square Errors)
Gradient Descent Optimizer	Adam
Batch Size	64
Holdout RMSE	32.47

TABLE 6. Seq-2-Seq architecture with 2 hidden layers in the encoder architecture.

Hidden Layers	Neurons
Hidden Layer 1(Encoder)	256
Hidden Layer 2(Encoder)	64
Hidden Layer 1(Decoder)	16
Output Layer	1
Loss Function	L2(Least Square Errors)
Gradient Descent Optimizer	Adam
Batch Size	64
Holdout RMSE	22.30

D. SEQUENCE-2-SEQUENCE NETWORK ARCHITECTURE

As mentioned in the previous section, Seq2Seq is a category of RNN/LSTMs which are mostly used for natural language processing and time series applications. Fig.2 shows a representational structure of a Sequence-2-Sequence model.

Similar to the approach of finding an optimal structure for a BiLSTM model, we also start out with single hidden layer in the encoder architecture of a Sequence-to-sequence model. The Tables 5, and 6 describe the different architectures:

For the Sequence-2-Sequence architecture, the addition of an extra hidden layer in the encoder architecture, leads to better encapsulation of the input features and therefore gives much better accuracy. Since, the encoder structure is important in an encoder-decoder architecture for it being crucial to encode important input features, the number of hidden layers only was investigated with having different number of layers.

The batch size was not decreased in this case of a Sequence-2-Sequence model as the models already gives a much better RMSE score for the test set. A batch size of 64 is used.

The standard optimizer, Adam, is again used with its default values.

At every layer, a Dropout of 20% is applied.

TABLE 7. Proposed bidirectional LSTM model architecture and design parameters for band forecast of DA/RT price differences with 1 hidden layer.

Hidden Layers	Neurons (forward & reversed sequence)
Hidden Layer	64 + 64
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	32
Test Accuracy	71.37%

TABLE 8. Proposed bidirectional LSTM model architecture and design parameters for band forecast of DA/RT price differences with 2 hidden layers.

Hidden Layers	Neurons (forward & reversed sequence)
Hidden Layer	64 + 64
Hidden Layer	32 + 32
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	64
Test Accuracy	87.11%

E. BAND FORECAST OF DA/RT PRICE DIFFERENCE

While value forecasting of the price differences is most desirable, forecasting the band/range of price differences is also valuable in some applications because it can increase the confidence level of the forecast. Therefore, in this work, we also apply the proposed Bidirectional LSTM and Sequence-to-Sequence networks to do band forecast. The bands of the price differences are defined as follows: there are in total 42 different price bands. The size of each price difference band varies by the value of price differences. The higher the value, the larger size of the band. Specifically, the size of the bands is $5/MWh for price difference between $-80/MWh and $80/MWh. Other bands are defined as (80, 100], (100, 150], (150, 250], (250, 350], (350, 500], (500, 700], (700, 1000], [-100, -80), [-150, -100), [-250, -150), all in the unit of $/MWh. The bands cover the complete range of price differences, including both positive and negative values.

In order to do band/range forecast, we try different architectures of the Bidirectional LSTM and the Sequence-2-Sequence models since the overall approach remains the same, except that the activation to the last layer (output layer) is changed to a *softmax* activation function to spit out the probabilities of the different categories.

Similar to the way we conducted the experiments for price difference value predictions, Tables 7, 8, 9 describe the architectures of the Bidirectional LSTM model used for classifying the price difference ranges. Since the same data is used for constructing the BiLSTM and Sequence-2-Sequence Networks, the same model architectures are used.

As we observe from the above tables, addition of hidden layers increases the overall test accuracy. This is again similar to our observation from our approach for tuning the architecture for the value prediction.

TABLE 9. Proposed bidirectional LSTM model architecture and design parameters for band forecast of DA/RT price differences with 3 hidden layers.

Hidden Layers	Neurons (forward & reversed sequence)
Hidden Layer	96 + 96
Hidden Layer	48 + 48
Hidden Layer	10 + 10
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	64
Test Accuracy	90.24%

TABLE 10. Proposed Seq2Seq model architecture and design parameters for band forecast of DA/RT price difference with 1 encoder hidden layer.

Hidden Layers	Neurons
Hidden Layer 1(Encoder)	50
Hidden Layer 1(Decoder)	20
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	64
Test Accuracy	92.17%

TABLE 11. Proposed Seq2Seq model architecture and design parameters for band forecast of DA/RT price difference with 2 encoder hidden layers.

Hidden Layers	Neurons
Hidden Layer 1(Encoder)	96
Hidden Layer 2(Encoder)	64
Hidden Layer 1(Decoder)	20
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	64
Test Accuracy	88.13%

TABLE 12. Proposed Seq2Seq model architecture and design parameters for band forecast of DA/RT price difference with 3 encoder hidden layers.

Hidden Layers	Neurons
Hidden Layer 1(Encoder)	128
Hidden Layer 2(Encoder)	64
Hidden Layer 3(Encoder)	32
Hidden Layer 1(Decoder)	20
Output Layer	42
Loss Function	Categorical Cross-entropy
Gradient Descent Optimizer	Adam
Batch Size	64
Test Accuracy	87.95%

Again, for the Sequence-2-Sequence band prediction, we use the same approach to find the most suitable architecture.

As we can observe from Tables 10, 11, and 12, increasing the number of hidden layers in the encoder structure, interestingly, decreases the accuracy of the model in the test set. Also, in such case, when we already have decreasing performance with increasing hidden layers in the encoder structure, increasing the decoder layers would not be of much help, since decoder only decodes from the encoded vector.

For these architectures, a dropout of 20% was applied to every layer for regularization. The weight matrices were initialized with Lecunn weights, and a kernel regularization of 0.001 was applied to all the layers.

III. COMPARISON AND EVALUATION METHODS

A. COMPARISON METHODS FOR VALUE FORECAST AND BAND FORECAST

In order to evaluate the value forecast performance of the proposed Bi-directional LSTM and Sequence-to-Sequence Networks, their performances are compared against those of a few popular methods such as Auto-Regressive Integrated Moving Average (ARIMA), Gradient Boosted Decision Trees (XGBoost), and Support Vector Regression (SVR), and a Random Walk Model. For band forecast performance of the proposed classification models based on Bi-directional LSTM and Sequence-to-Sequence Networks, it is compared with that of Multinomial Logistic Regression, Support Vector Classifier (SVC), and XGB Classifier. The comparison methods are briefly introduced as follows.

ARIMA is considered as the "gold-standard" among methods which have been used to forecast market clearing prices reliably and accurately [26]. Generally, an ARIMA model is fitted to a time series by some statistical software like R. These models have 3 parameters which can be determined manually by observing the autocorrelation plots, or can also be calculated automatically by the software. The 3 parameters include the time lag values for the Auto-regressive part (p), the degree of differencing (d), and that for the Moving Average part (q).

The automatic feature of R of determining the parameter values p, d, q is used. The values determined after fitting the model are:

p = 5, d = 0, q = 3.

Gradient-boosted Decision Trees These are decision trees, where gradient descent is used to minimize the loss when adding multiple decision trees to the model using boosting. [27] uses XGBoost model for price forecasting, with good accuracy.

Support Vector Regression (SVR)/Support Vector Classification (SVC) is a very common machine learning algorithm for solving regression, or classification problems, respectively. Authors in [28] use SVR technique to forecast prices in electricity markets.

Multinomial Logistic Regression It is a classification method that generalizes logistic regression to multiclass problems. This type of classifier is used to predict the probabilities of the different possible outcomes of a dependent variable distributed based on different categories, given a set of independent variables.

Random Walk Model This model predicts that the future value, k timesteps ahead, will be equal to the present value. Mathematically, it can be expressed as:

$$\hat{Y}_{n+k} = Y_n,$$

where $\hat{Y}_{n+k}$ is the forecast value k timesteps into the future and Y_n is the present value.

B. PERFORMANCE EVALUATION OF VALUE FORECAST

For value forecasting of prices, Root Mean Squared Error (RMSE) values are used to measure the models' performance. RMSE values in the case study is in the unit of $/MWh.

Also, zoomed-in snapshots of forecasting results are presented in order to validate the claim of the trained algorithms being able to capture the spikes in the correct direction and at the correct time.

For applications like virtual bidding, when it is difficult to predict the magnitude of the price difference, it is still valuable to predict the sign of the price difference. In order to capture the performance in this regard, a separate metric is devised. We define *Direction Accuracy* as follows:

$$Direction\_Accuracy = (\frac{n}{N} * 100)$$

where n = total number of correctly detected direction (or sign) of price difference, N = total number of price differences. Since the series are already time differenced, the value of n is determined by whether the predicted and actual values have the same sign, i.e., both have positive or negative sign at the same timestep.

For a good forecasting linear method, it is expected that it will produce forecasting residuals that do not show observed patterns. Therefore, we use two tests to determine the randomness of the residuals for the ARIMA model.

1) WALD-WOLFOWITZ RUNS TEST

The Wald–Wolfowitz runs test (or simply runs test), named after statisticians Abraham Wald and Jacob Wolfowitz is a statistical test that checks a randomness hypothesis for a two-valued data sequence [29]. For the Runs test, the null hypothesis is defined as:

H_o: The residuals follow a random distribution
H_a: The residuals do not follow a ransom distribution

2) SHAPIRO-WILK TEST FOR NORMALITY

The Shapiro–Wilk test is a test of normality in statistics. The null-hypothesis of this test is that the population is normally distributed. The definition of the null hypothesis is:

H_o: The residuals follow a normal distribution
H_a: The residuals do not follow a normal distribution

C. PERFORMANCE EVALUATION OF SPIKE VALUE FORECAST

For DA/RT price difference forecast, the ability of forecasting extreme values are of most importance as the timesteps with high price difference values indicate larger deviation between DA and RT market prices and higher opportunity for virtual bidding. Because the extreme values occur for only one or very few timesteps, it is particularly difficult to forecast their occurrence at the right timesteps. Missing any one of the

TABLE 13. Value forecast performance of different forecasting methods.

Model	RMSE ($/MWh)	Direction Accuracy (%)
ARIMA(5, 0, 3)	48.17	59.30
XGBoost	39.69	75.95
SVR	46.29	76.25
Random Walk	43.78	80.12
Bidirectional LSTM	33.15	75.54
Seq2Seq	22.30	79.41

TABLE 14. Statistical test for randomness and normality.

Model	p-value(Runs Test)	p-value(Shap.-W Test)
ARIMA	0.155	0.0502

extreme values will cause significant increase in Mean Absolute Percentage Error (MAPE). Therefore, we use MAPE to evaluate the performance of peak value forecast.

D. PERFORMANCE EVALUATION OF BAND FORECAST

For band forecast, the commonly used accuracy metric is used to evaluate the correctness of the proposed classification models. The accuracy percentage is calculated by summing up the total number of correctly predicted bands, and then dividing it by the total number of samples.

If the forecasted band is the same band as the one the actual price difference falls within, the forecast band is considered a correct forecast.

IV. RESULTS AND DISCUSSION

A. VALUE FORECAST RESULTS

The proposed Bidirectional LSTM and Seq-2-Seq models are implemented using Keras library with Tensorflow backend. The code is written in Python. The models are trained for 2000 epochs. Comparison models ARIMA, XGBoost and SVR were all tested on the same test set. For the SVR and XGBoost models, the time lag used is 48, the same as that was used in the Bidirectional LSTM model. The performance of different models in predicting the next hour's price value is shown in Table 13.

Moreover, Figs.3, 4 and 5 show in time series the DA/RT market price difference forecasting performance of the different models in three zoomed-in time windows of 2018. The time windows are selected to show the performance during low, medium, and high level of price differences respectively.

Table 13 shows the Bidirectional LSTM and Seq2Seq models achieved better accuracy in forecasting magnitude that is measured in RMSE. The proposed deep learning models also more accurately predicted the sign of the DA/RT price difference, measured in Direction Accuracy. When zooming into the forecast performance at each timestep, each model may perform well at some timesteps and not at other timesteps. Overall, the deep learning methods produced forecast values that are more in agreement with the actual values. More importantly, as seen in Figs.3, 4 and 5, the deep learning methods can better capture the spikes at the right timesteps. Between the two deep learning methods, the Seq2Seq model performed exceptionally well and even outperforms the Bidirectional LSTM model.

Although the direction accuracy value of the Random Walk model might tempt one into believing that a random

walk model might have some value, other metrics show the opposite: (1) the RMSE value of the random walk model is among the highest, showing that the model performs worse as compared to other models when taking into account how much off it is than the real value, (2) the Random Walk model always forecasts the value of the present timestep as the future value for the next timestep. Therefore, it cannot pick up any spike at the right timestep, since it is always off by 1 timestep in this case.

B. RESIDUAL ANALYSIS

In this subsection, the residuals of the ARIMA model is visualized and statistical tests results are presented.

From the residual plot shown in Fig.6 we can observe that for the tested model the residuals are mostly concentrated at the centre fit. Few points are scattered on either side of the plots. Moreover, no clear patterns are observed from the scattered points on either side of the lines in all the plots. In order to have a more conclusive evidence, the p-value calculated by Runs Test and Shapiro-Wilk Test are shown in Table 14.

As we can observe from Table 14, for Runs Test, the p-values for the ARIMA model is greater than 0.05. This indicates that, the residuals come from a random distribution with 95% confidence. This means that no patterns are found in the model. In other words, that the fitted ARIMA model presents a good fit to the data set.

For Shapiro-Wilk Test, the p-values for the model is above 0.05. It means that the residuals come from a normal distribution with 95% confidence.

### C. PERFORMANCE ON FORECASTING SPIKE VALUES

It should be noted that, in order to capture the magnitude of spike values, a model needs to be able to capture the timing of such spikes, which however does not present clear cyclic pattern. It is therefore quite difficult to capture the magnitudes. For extremely high values over $100/MWh values, Seq2Seq model can achieve 20% MAPE. Bidirectional LSTM has a MAPE of 54%, capturing about half magnitude of the extreme values. In contrast, ARIMA, XGBoost and SVR missed most or the complete magnitudes of the extreme values with 78%-100% MAPE, signifying the difficulty of forecasting peak values at the right timesteps. The Random Walk model is not able to capture any spike at the correct timestep since it always predicts the correct value 1 timestep later, and hence not considered in the Spike Performance analysis. Detailed performance results in MAPE are shown in Table 15.

It should be pointed out that, some models such as XGBoost can forecast peak value with some time lag
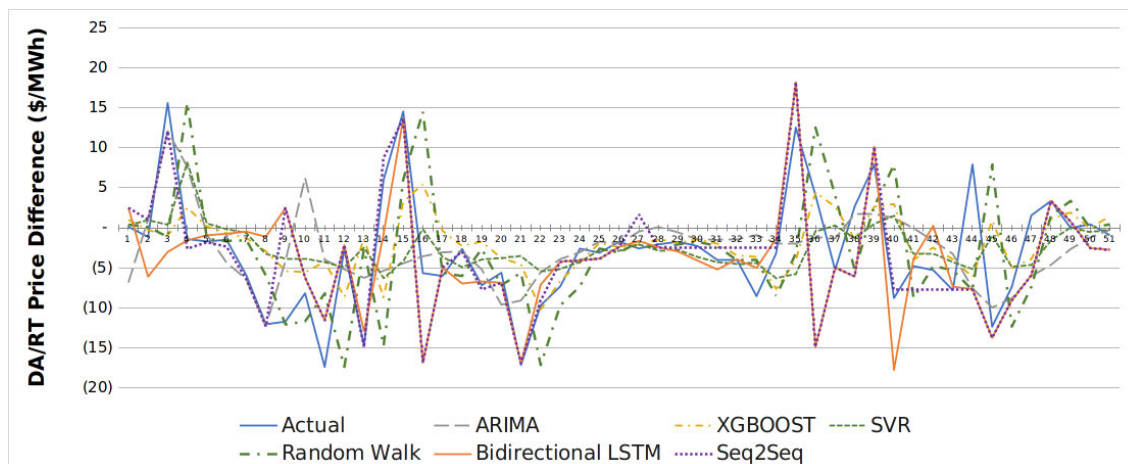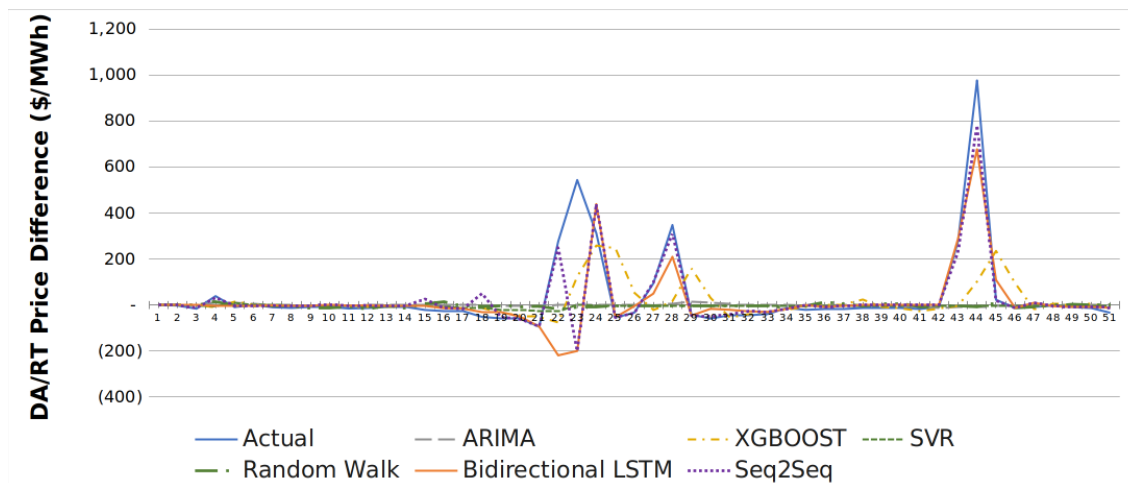
FIGURE 3. Actual and forecasted DA/RT market price difference using PJM market data from Oct. 31, 2018 11:00 PM through Nov. 3, 2018 1:00 AM.
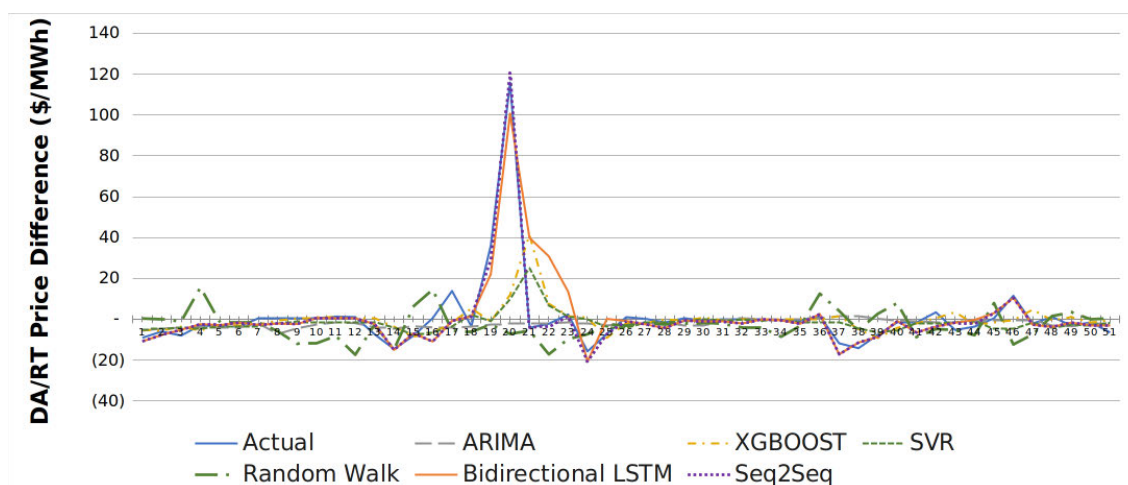
FIGURE 4. Actual and forecasted DA/RT market price difference using PJM market data from Nov. 11, 2018 6:00 PM through Nov. 13, 2018 8:00 PM.

FIGURE 5. Actual and forecasted DA/RT market price difference using PJM market data from Dec 2, 2018 2:00 PM through Dec 4, 2018 4:00 PM.
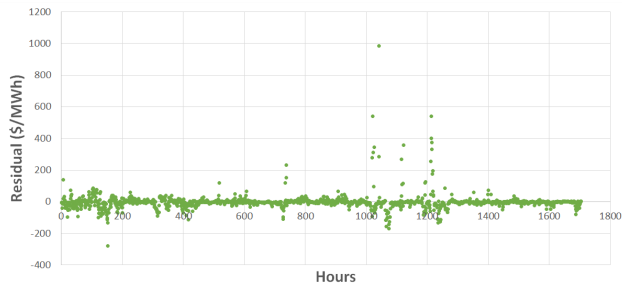
FIGURE 6. Residual Plot of ARIMA model for October 22, 2018 through December 31, 2018.

TABLE 15. Spike value forecast performance in MAPE.

Model	MAPE (%) For Timesteps with Actual Price Diff.$\geq$100 $/MWh	MAPE (%) For Timesteps with Actual Price Diff.$\geq$50 $/MWh
ARIMA	95	92
XGBoost	69	79
SVR	99	97
Random Walk	59	68
Bi-directional LSTM	53	62
Seq2Seq	19	30

TABLE 16. Spike value forecast performance in direction accuracy.

Model	Direction Accuracy (%) For Timesteps with Actual Price Diff.$\geq$100 $/MWh	Direction Accuracy (%) For Timesteps with Actual Price Diff.$\geq$50 $/MWh
ARIMA	63	70
XGBoost	84	78
SVR	53	65
Random Walk	84	81
Bi-directional LSTM	87	84
Seq2Seq	95	93

(often one timestep lag), as seen in Fig.5. This slight time lag may be acceptable in some applications, but can be devastating in applications like virtual bidding, and may lead to wrong decisions and undesired financial outcome.

For actual price difference above $50/MWh, Seq2Seq still achieved 30% MAPE, the best among the five different models. In other words, roughly speaking the Seq2Seq model successfully captured 70% of the magnitudes of spike values that are higher than $50/MWh$.

As shown in Table 16, for timesteps with high values of DA/RT price difference, Seq2Seq model achieved an exceptionally good direction accuracy of over 90%. It means the Seq2Seq model is able to accurately capture the sign of the DA/RT price difference for vast majority of those timesteps, which provides great value to market participants.

D. SELECTING MODELS USING THE MODEL CONFIDENCE SET

The Model Confidence Set, proposed by [30], is used to select a subset of ideal candidate models among all other forecasting models based on a confidence level, α.

TABLE 17. Model confidence Set p-values.

Model	MCS p-value
ARIMA	0.0482
XGBoost	0.0892
SVR	0.012792
Random Walk	0.10773
Bi-directional LSTM	0.14938
Sequence-2-Sequence	0.3917

The *estMCS* package in R was used in order to estimate the MCS p-values for the different models. The MCS p-values, where $p_i \geq \alpha$, are generally considered in the subset of the ideal models. The *loss* function used to compare the different models is the Root Mean Squared Error (RMSE) for all the models.

Those models which have the minimum expected RMSE, are defined as the best ones. The Model Confidence Set works to, first, remove inferior candidates. It is done so by conducting hypothesis tests in a sequential manner. The Null and alternate hypothesis statements are given by:

Null Hypothesis: No inferior model is present

Alternate Hypothesis: At least one inferior model is present

If the p-value is less than the significance level chosen, the null hypothesis is rejected, the model is removed from the confidence set, and the null hypothesis is tested again. Also, when the test statistic fails to reject the null hypothesis, the overall procedure of model elimination stops, and the remaining models are given by the algorithm as the final models in the Model Confidence Set.

As we observe in Table.17, at a significance level, $\alpha = 0.05$, we have a set of the following models: {Sequence-2-Sequence, Bi-directional LSTM, XGBoost, and Random Walk}. This says that we have this set of ideal models for forecasting at 95% confidence.
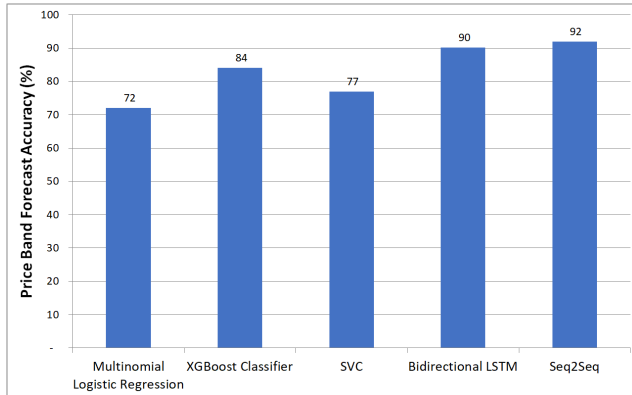
From the Table 17, we see that our final set based on the pre-defined significance level results in the following 4 models: Sequence-2-Sequence, Bi-directional LSTM, XGBoost, and Random Walk. However, since Random Walk model can not detect spikes at the correct time step and always predicts the spike after 1 time step, this method is not suitable for our purpose.

E. BAND FORECAST RESULTS

In order to validate deep learning methods' capability to forecast the bands of price differences, the same deep learning methods including the Bidirectional LSTM and the Sequence-to-Sequence networks are used as multi-class classifiers. As shown in Fig.7, the deep learning models achieved significantly higher accuracy in band forecast than other common methods such as Multinomial Logistic Regression, XGBoost Classifier and SVC. Sequence-to-Sequence network consistently performed the best among all tested models in the classification task.

F. TEST GENERALIZATION ABILITY OF Seq2Seq MODEL

As mentioned previously, the deep learning and other machine learning models were trained on the data from

FIGURE 7. Accuracy of band forecast of different models.

FIGURE 8. Forecasting performance of Seq2Seq on data from 2019 for Feb. 20, 2019 3:00 AM through Feb. 25, 2019 9:00 PM.

TABLE 18. Value forecast performance of Seq2Seq model for January-February 2019 of PJM market.

Model	RMSE($/MWh)	Direction Accuracy(%)
Seq2Seq	19.47	73.28

2018. However, an interesting question arises: whether the proposed deep learning models can be generalized over time. An important and desirable feature about deep learning is whether a model can be used, without re-training it, on another dataset and give good forecasting results. Therefore, we apply the deep learning models which are trained using the first 9 months of 2018 PJM data, to test the forecasting performance for January to end of February of year 2019. In the forecasting test, the model was not retrained using 2019 dataset. The previously trained model was used directly to forecast the hourly LMP difference in 2019. The forecasting performance using the evaluation metrics is summarized in Fig.8 and Table 18. Results show the Seq2Seq model performs surprisingly well on the new dataset without being retrained.

Despite the generalization ability over time, it is in our opinion that the generalization ability over price locations is expected to be limited, except for nearby locations. It is because some major factors impacting the price deviation between DA and RT markets such as real-time outages and the resulting change in congestion have strong impact on local prices and weaker impact on prices at remote locations. Therefore, the implicit pattern seen at one location may be quite different from that at another location.

G. SUMMARY

Various performance metrics have been used to compare between the proposed deep learning models and other statistical and machine learning methods, and to evaluate the performance in predicting values, its direction, magnitude, timing as well as the performance in predicting bands.

From Table 13, we can see Sequence-to-Sequence model produces significantly smaller RMSE than other methods, and also the Sequence-to-Sequence model gives the highest accuracy in forecasting the direction of DA/RT market price differences.

Moreover, a careful investigation of Fig.3, Fig.4 and Fig.5 reveals that the Sequence-to-Sequence approach is able to exceptionally well predict the occurrences of spikes of DA/RT price differences, which conventional methods
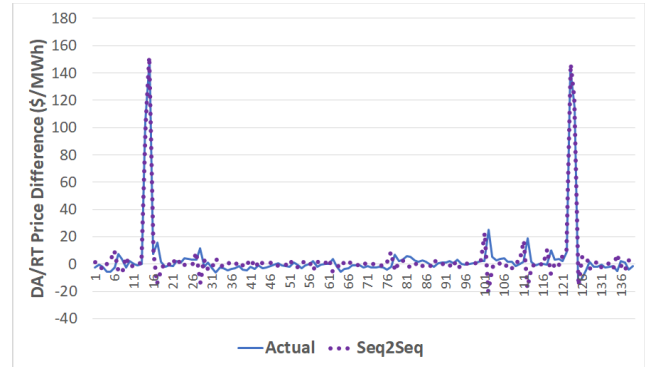
find difficult. In order to verify whether the Sequence-to-Sequence model can generalize over time, the proposed model is tested on the first 3 months of 2019, and the performance is still very good.

The performance in forecasting the spikes values is quantified in MAPE. Table 15 and Table 16 show the Seq2Seq model can capture 81% of the magnitude and 95% of the direction of the price difference values that are greater than $100/MWh.

The multi-class classifiers constructed using the proposed deep learning models also achieved over 90% accuracy in forecasting the price difference band.

Finally, for applications like virtual bidding in the DA market, a model capable of predicting the value and direction of the DA/RT price difference, and magnitude and timing of spike values, is of utmost importance. Our proposed deep learning models, especially the Seq2Seq model, provide a promising solution.

V. CONCLUSION AND FUTURE WORK

This paper proposed deep, Bidirectional Long-Short Term Memory and Sequence-to-Sequence architectures to forecast nodal LMP difference between the DA and RT markets. The models' forecasting performances are measured using the Test RMSE values, and also tested on a different year's data which prove the generalization capability of the models. The models' capabilities in correctly forecasting spikes, directions and bands of DA/RT price difference at the right timesteps are also demonstrated using the custom metric and the classification models. The deep learning models are compared with ARIMA, XGBoost, SVR and Random Walk methods and outperform in all tested evaluation metrics. Moreover, the Sequence-to-Sequence architecture outperforms even the Bidirectional LSTM model.

In addition to DA/RT market price difference forecast, the proposed approach has the potential to be applied to solve

other forecasting problems such as price spread forecast in DA market for Financial Transmission Right (FTR) trading purpose.

Future research may include expanding this approach to multivariate time series forecasting, multi-hour price difference forecasting, and improving generalization ability. First, this paper only deals with a univariate time series model, where only the time series price differences are considered. The prediction may be further improved by taking into account other influencing factors such as load and Day-ahead Market LMP. Second, the proposed model currently only forecasts price difference for the next hour. It should be expanded to multihour (such as 24 hours in US electricity markets) to align with market practices. Lastly, the architectures of the deep learning methods will need to be modified for the forecasting problem at different nodes or in different markets. It will be of great interest to improve the generalization ability of the deep learning methods.

ACKNOWLEDGMENT

REFERENCES

[1] M. Shafie-khah, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "Price forecasting of day-ahead electricity markets using a hybrid forecast method," *Energy Convers. Manage.*, vol. 52, pp. 2165–2169, May 2011.

[2] D. Singhal and K. S. Swarup, "Electricity price forecasting using artificial neural networks," *Int. J. Elect. Power Energy Syst.*, vol. 33, no. 3, pp. 550–555, 2011.

[3] A. J. Conejo, J. Contreras, R. Espínola, and M. A. Plazas, "Forecasting electricity prices for a day-ahead pool-based electric energy market," *Int. J. Forecasting*, vol. 21, no. 3, pp. 435–462, Jul. 2005.

[4] Y. Ji, R. J. Thomas, and L. Tong, "Probabilistic forecasting of real-time LMP and network congestion," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 831–841, Mar. 2017.

[5] Y.-Y. Hong and C.-P. Wu, "Day-ahead electricity price forecasting using a hybrid principal component analysis network," *Energies*, vol. 5, no. 11, pp. 4711–4725, Nov. 2012.

[6] R. Bo and F. Li, "Probabilistic LMP forecasting considering load uncertainty," *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1279–1289, Aug. 2009.

[7] S. S. Reddy, C.-M. Jung, and K. J. Seog, "Day-ahead electricity price forecasting using back propagation neural networks and weighted least square technique," *Frontiers Energy*, vol. 10, no. 1, pp. 105–113, Mar. 2016.

[8] A. R. Gollou and N. Ghadimi, "A new feature selection and hybrid forecast engine for day-ahead price forecasting of electricity markets," *J. Intell. Fuzzy Syst.*, vol. 32, no. 6, pp. 4031–4045, May 2017.

[9] R. K. Agrawal, F. Muchahary, and M. M. Tripathi, "Ensemble of relevance vector machines and boosted trees for electricity price forecasting," *Appl. Energy*, vol. 250, pp. 540–548, Sep. 2019.

[10] Y.-Y. Hong, J. V. Taylar, and A. C. Fajardo, "Locational marginal price forecasting in a day-ahead power market using spatiotemporal deep learning network," *Sustain. Energy, Grids Netw.*, vol. 24, Dec. 2020, Art. no. 100406.

[11] K. Zheng, Y. Wang, K. Liu, and Q. Chen, "Locational marginal price forecasting: A componential and ensemble approach," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4555–4564, Sep. 2020.

[12] X. Ma, D. I. Sun, G. W. Rosenwald, and A. L. Ott, "Advanced financial transmission rights in the PJM market," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, Jul. 2003, pp. 1031–1038.

[13] L. Hadsell, "The impact of virtual bidding on price volatility in New York's wholesale electricity market," *Econ. Lett.*, vol. 95, no. 1, pp. 66–72, Apr. 2007.

[14] S. Baltaoglu, L. Tong, and Q. Zhao, "Algorithmic bidding for virtual trading in electricity markets," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 535–543, Jan. 2019.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] R. C. Garcia, J. Contreras, M. V. Akkeren, and J. B. C. Garcia, "A GARCH forecasting model to predict day-ahead electricity prices," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 867–874, May 2005.

[17] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.

[18] X. Hu, D. V. Prokhorov, and D. C. Wunsch, "Time series prediction with a weighted bidirectional multi-stream extended Kalman filter," *Neurocomputing*, vol. 70, nos. 13–15, pp. 2392–2399, 2007.

[19] L. Yang and B. Zou, "A model research on PJM day-ahead market locational marginal price spread," in *Proc. IEEE 7th Int. Conf. Smart Energy Grid Eng. (SEGE)*, Aug. 2019, pp. 283–289.

[20] R. Das, R. Bo, W. Ur Rehman, H. Chen, and D. Wunsch, "Cross-market price difference forecast using deep learning for electricity markets," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Eur.)*, Oct. 2020, pp. 854–858.

[21] T. Teräsvirta, C.-F. Lin, and C. W. J. Granger, "Power of the neural network linearity test," *J. Time Ser. Anal.*, vol. 14, no. 2, pp. 209–220, Mar. 1993.

[22] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 3285–3292.

[23] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

[24] Y. Xie, K. Zhao, Y. Sun, and D. Chen, "Gaussian processes for short-term traffic volume forecasting," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2165, no. 1, pp. 69–78, Jan. 2010.

[25] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[26] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1014–1020, Aug. 2003.

[27] P. Li and J.-S. Zhang, "A new hybrid method for China's energy supply security forecasting based on ARIMA and XGBoost," *Energies*, vol. 11, no. 7, p. 1687, Jun. 2018.

[28] J. Che and J. Wang, "Short-term electricity prices forecasting based on support vector regression and Auto-regressive integrated moving average modeling," *Energy Convers. Manage.*, vol. 51, pp. 1911–1917, Oct. 2010.

[29] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, no. 4, pp. 697–717, Jul. 1979.

[30] P. R. Hansen, A. Lunde, and J. M. Nason, "The model confidence set," *Econometrica*, vol. 79, no. 2, pp. 453–497, 2011.

• • •