Konya Mühendislik Bilimleri Dergisi, c. 11, s. 4, 1047-1060, 2023
Konya Journal of Engineering Sciences, v. 11, n. 4, 1047-1060, 2023
ISSN: 2667-8055 (Electronic)
DOI: 10.36306/konjes.1290652

A TWO STAGE MODEL FOR DAY-AHEAD ELECTRICITY PRICE FORECASTING: INTEGRATING EMPIRICAL MODE DECOMPOSITION AND CATBOOST ALGORITHM

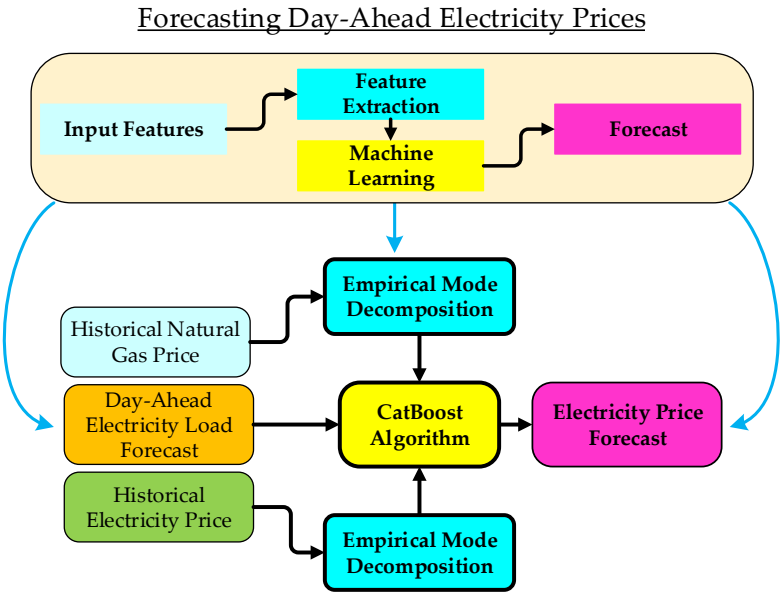
Ceyhun YILDIZ

Kahramanmaraş İstiklal University, Elbistan Vocational School, Electricity and Energy Department, Kahramanmaraş, TÜRKİYE
ceyhun.yildiz@istiklal.edu.tr

Highlights

- A two-stage model for day-ahead electricity price forecasting was introduced
- The proposed model combines empirical mode decomposition and CatBoost algorithm
- Empirical mode decomposition was employed for feature extraction
- CatBoost algorithm was utilized for electricity price forecasting
- Comparisons with benchmark models comfirmed the effectiveness of the proposed model

Graphical Abstract



Proposed electiricty price forecasting model

***Corresponding Author:** Ceyhun YILDIZ, ceyhun.yildiz@istiklal.edu.tr

Konya Mühendislik Bilimleri Dergisi, c. 11, s. 4, 1047-1060, 2023
Konya Journal of Engineering Sciences, v. 11, n. 4, 1047-1060, 2023
ISSN: 2667-8055 (Electronic)
DOI: 10.36306/konjes.1290652

A TWO STAGE MODEL FOR DAY-AHEAD ELECTRICITY PRICE FORECASTING: INTEGRATING EMPIRICAL MODE DECOMPOSITION AND CATBOOST ALGORITHM

Ceyhun YILDIZ

Kahramanmaraş İstiklal University, Elbistan Vocational School, Electricity and Energy Department, Kahramanmaraş, TÜRKİYE
ceyhun.yildiz@istiklal.edu.tr

(Received: 01.05.2023; Accepted in Revised Form: 03.10.2023)

ABSTRACT: Electricity price forecasting is crucial for the secure and cost-effective operation of electrical power systems. However, the uncertain and volatile nature of electricity prices makes the electricity price forecasting process more challenging. In this study, a two-stage forecasting model was proposed in order to accurately predict day-ahead electricity prices. Historical natural gas prices, electricity load forecasts, and historical electricity price values were used as the forecasting model inputs. The historical electricity and natural gas price data were decomposed in the first stage to extract more deep features. The empirical mode decomposition (EMD) algorithm was employed for the efficient decomposition process. In the second stage, the categorical boosting (CatBoost) algorithm was proposed to forecast day-ahead electricity prices accurately. To validate the effectiveness of the proposed forecasting model, a case study was conducted using the dataset from the Turkish electricity market. The proposed model results were compared with benchmark machine learning algorithms. The results of this study indicated that the proposed model outperformed the benchmark models with the lowest root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (R) values of 8.3282%, 5.2210%, 6.9675%, and 86.2256%, respectively.

Keywords: *CatBoost regression, Electricity price forecasting, Empirical mode decomposition*

1. INTRODUCTION

Forecasting electricity prices is essential for the secure and economic operation of interconnected electrical grid systems. Power generation companies use price forecasts to plan their generation, aiming to maximize their profit, while consumers use forecasts to avoid high electricity prices and minimize their costs. However, with the liberalization of the energy industry, electricity price data presents increasingly complex dynamics and uncertainties. The liberalized electricity market operations are highly intricate due to stochastic factors such as meteorological conditions, the balance of generation and demand, grid system constraints, fuel prices, and energy policies. As a result of the complex relationships between these factors, predicting electricity prices with a high level of accuracy is challenging.

Over the past twenty years, numerous approaches and models have been proposed for electricity price forecasting. Comprehensive literature reviews on the study of electricity price forecasting were introduced by [1] and [2]. The approaches in the literature for predicting electricity prices can be divided into three main sub-sections: (1) statistical models, (2) artificial intelligence models, and (3) hybrid models.

Statistical models typically use historical data, as well as other input features, to perform the forecasting process. Although statistical models were frequently employed for predicting electricity prices, they faced some restrictions and difficulties. Unlike the linear relationships between variables that statistical models are intended to model, typical electricity price forecasting models capture the non-linear dynamics of electricity markets. The significance of taking nonlinearities into consideration when forecasting electricity prices was pointed out in the review article [2], which also discussed different approaches to this issue.

Artificial intelligence-based electricity price forecasting models employ machine learning (ML) and specifically deep learning (DL) approaches. In [3], different ML algorithms were investigated for day-

***Corresponding Author:** Ceyhun YILDIZ, ceyhun.yildiz@istiklal.edu.tr

ahead electricity price forecasting, and electricity market datasets from European countries were used to evaluate forecasting algorithms. An online self-adaptive forecasting method based on random forest (RF) was proposed in [4] to forecast electricity prices, which considers the concept drift phenomenon of the power market. The case studies were conducted on the data from Gansu province, China, to illustrate its effectiveness. An interesting approach based on an online sequential extreme learning machine (OS-ELM) was proposed in [5] for forecasting day-ahead and real-time market electricity prices. Performance tests were conducted using the dataset from the Australian Energy Market Operator (AEMO). In [6], deterministic and stochastic components of the electricity price signal were investigated, and several parametric and nonparametric approaches were employed for deterministic and stochastic component forecasting. The experimental studies were carried out using the dataset from the Italian electricity market (IPEX). An interesting DL model was proposed in [7] to forecast electricity prices. The proposed model was the new deep convolutional neural network (CNN) architecture based on GoogLeNet. The effectiveness of this model was tested on the dataset from the New York Independent System Operator (NYISO). Another model for NYISO data was introduced in [8]. The authors proposed a CNN-based autoencoder to forecast electricity prices. In [9], four DL-based electricity price forecasting models were proposed, and the results were compared with 27 state-of-the-art predictors. Simple neural network (NN) and DL-based models for electricity price forecasting were introduced in [10], and the experimental studies were conducted using the dataset from the Electric Reliability Council of Texas (ERCOT). However, the main drawbacks of these deep networks are their computational expense and black-box nature.

The third literature section comprises hybrid model approaches, several data preprocessing techniques, artificial intelligence algorithms, and statistical models that were combined to forecast electricity prices. [11] proposed a hybrid model based on data decomposition and the extreme learning machine (ELM) algorithm. The differential evolution (DE) algorithm was used to optimize ELM parameters. Electricity price datasets from Spanish and Australian electricity markets were used in the experimental study. [12] introduced a hybrid model for electricity price forecasting. The empirical wavelet transform (EWT) and an attention mechanism were proposed to decompose and select input features, respectively. The long-short-term-memory (LSTM) architecture was used to obtain final forecasts. Furthermore, the crisscross optimization algorithm (CSO) was used to determine fully connected layer parameters. Wind power generation, solar power generation, predicted electricity load, and historical electricity price data from the Danish energy market were used as model inputs. Another hybrid method based on the wavelet transform (WT) and an LSTM was proposed in [13] for forecasting electricity prices. The data from the Pennsylvania, New Jersey, and Maryland (PJM) and Spain electricity markets were used to evaluate the proposed method. A hybrid method based on model input decomposition and forecasting was proposed in [14]. EMD and DL-based methods were employed for decomposition and forecasting, respectively. The performance evaluations were realized on the datasets from the PJM and New South Wales electricity markets. [15] introduced an interesting hybrid model for day-ahead electricity price forecasting. The authors proposed an adaptive copula-based method and a new signal decomposition algorithm for feature selection and extraction. The effectiveness of the proposed model was validated using the dataset from the PJM electricity market. [16] proposed a two-stage ensemble learning-based approach for electricity price forecasting. The first stage comprises extreme gradient-boosted trees (XGBoost) and random forest (RF) to learn distinct features of the electricity price signal. The second stage includes Bayesian linear regression to obtain the final forecast values. The proposed model was tested on the Austrian electricity market dataset. [17] proposed another hybrid model based on CNN and LSTM. CNN was employed for extracting new features from model inputs. The extracted features were used as LSTM model inputs to obtain electricity price forecasts. The dataset from the Iranian electricity market was used to train and test the proposed forecasting model. Another approach was proposed in [18]. The authors proposed data augmentation methods and regression models (multilayer neural network (MLNN), CNN, and autoregressive model with exogenous inputs (ARX)) to forecast day ahead electricity prices. Electricity price datasets from Belgian and Dutch day-ahead electricity markets were employed in

the study. [19] proposed an electricity price forecasting model that combines WT, stacked autoencoder (SAE), and LSTM for the U.S. energy market. [20] introduced a hybrid model based on decomposition and forecasting. WT and LSTM were employed for decomposition and forecasting processes. The data set from the Australian Energy Market Operator (AEMO) was used for the experimental study. Another hybrid model was developed in [21] to forecast electricity prices in two main stages. An autoregressive time varying (ARXTV) model with exogenous variables was employed in the first stage to forecast electricity price values. In the second stage, support vector machine (SVM) and kernel regression models were used to detect and estimate price spikes. Another two-stage model based on feature selection and regression was proposed in [22]. The multi-objective binary-valued backtracking search algorithm (MOBBSA) and an optimized adaptive neuro-fuzzy inference system (ANFIS) were employed for feature selection and regression. The dataset from the Ontario power market was used to develop and test the proposed model. [23] proposed a hybrid model for electricity price forecasting that combines the artificial neural network (ANN) and the artificial cooperative search algorithm (ACS). Dataset from the Ontario electricity market was used in the study. [24] proposed another ensemble learning based model for electricity price forecasting in IPEX. The proposed model forecasts deterministic and stochastic components of the electricity price signal with semi-parametric techniques and ML algorithms, respectively. Another method that combines linear regression automatic relevance determination (ARD) and ensemble bagging extra tree regression (ETR) models was proposed in [25]. The experimental test studies of the study were conducted on a dataset from the Nord Pool electricity market. The LSTM and signal decomposition based model was proposed in [26]. The electricity price signal was decomposed and the tuned LSTM architectures were used to forecast high and low-frequency components. The sequence model-based optimization (SMBO) was employed for hyper parameter tuning. The datasets from the electricity markets of PJM were used in this study. [27] introduced another hybrid model based on DL architectures. The proposed model combines deep belief network (DBN), CNN, and LSTM for feature extraction and regression processes. The PJM market data was used in this study.

The papers reviewed above indicated that research in the field of electricity price forecasting has largely focused on the hybrid approach. Despite the impressive progress made in the literature on hybrid machine learning models, there are still a number of gaps that need to be filled. Many of the proposed hybrid electricity price forecasting models have a black-box nature, making it difficult to interpret how they make predictions. This is a significant issue, as it's important to have explainable models. Additionally, some hybrid models are computationally expensive, which limits their scalability to large datasets. Therefore, more research is needed to develop hybrid models that are both accurate, explainable, and computationally efficient.

This study proposes a hybrid model to accurately forecast day-ahead electricity prices. The model architecture comprises two main stages. In the first stage, data preprocessing, feature selection, and feature extraction processes were performed to determine appropriate input features. The second stage employed a computationally efficient and explainable regression method to obtain reliable and accurate forecasts. Additionally, several performance tests were conducted to confirm the effectiveness of the proposed forecasting model architecture.

This study makes the following contributions:

1-Introducing a hybrid method based on feature selection, feature extraction, and regression.

2-Employing the mutual information (MI)-based feature selection method to determine the important features of the investigated dataset. Then, using an effective signal decomposition algorithm called EMD to extract meaningful deep features.

3- Proposing the computationally efficient and powerful algorithm CatBoost for regression.

4- Conducting performance comparison tests with benchmark regression algorithms on the dataset from the Turkish electricity market.

The rest of the paper is organized as follows: The upcoming section describes the theoretical foundation of the methods used in this study and the structure of the proposed electricity forecasting model. Afterward, the section titled Data Description presents information on the datasets utilized in the

study. Then, the EMD Based Feature Extraction section describes the feature extraction procedure utilized in the study. The section called Experimental Study gives the details of the experiments conducted. The Results section presents a summary of the results. Finally, the last section concludes the study.

2. THEORETICAL BACKGROUND

In this study, a two-stage hybrid forecasting model was developed for day-ahead electricity price forecasting. The proposed model performs feature selection, feature extraction, and regression processes to forecast electricity prices using several effective methods. The first stage of the model employs MI-based feature selection and EMD-based feature extraction methods. In the second stage, the computationally efficient and explainable CatBoost algorithm was used for regression. The following subsections provide an overview of the theoretical basis of the methods employed in the study.

2.1. Mutual Information

Mutual information (MI) was first introduced as a measure of uncertainty in [28], to quantify the amount of information that can be transmitted between two systems. Since then, it has become widely used as a statistical measure of the dependence between two random variables in various fields. MI measures the information shared between variables, with a high MI indicating a strong relationship and a low MI indicating a weak relationship. The following equation defines the MI between two discrete random variables.

$$MI(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{1}$$

where X and Y are random variables with the joint distribution $P(x,y)$. $P(x)$ and $P(y)$ are the marginal distributions.

In this study, MI was used as a criterion for feature selection. The dataset investigated includes multiple features that may have a relationship with electricity prices. To determine the importance of each feature, MI values were calculated for each feature. Based on the calculated MI values, the two most important features, load forecast, and natural gas price were determined as exogenous inputs for the proposed electricity price forecasting model.

2.2. Empirical Mode Decomposition

The EMD algorithm was introduced in [29] for decomposing nonlinear and non-stationary signals. This algorithm has been widely used in various fields for extracting and analyzing the deep components of complex signals. The EMD algorithm performs a data-driven method that does not require any predefined mathematical model for the signal and can adaptively decompose the signal. This algorithm decomposes any complex signal into its simpler intrinsic mode functions (IMFs) and a residue. An IMF is an intrinsic oscillatory component of the original signal, and the residue represents the trend. An IMF was defined in [29] as a function with the following two properties: 1) the total number of local minimums and local maximums, and the number of zero crossings differ by at most one, and 2) the upper and lower envelopes derived from the local extrema have a mean value of zero. The following equation defines the relationships between the original signal and the extracted signals (IMFs and residue).

$$x(t) = \sum_{i=1}^{N} IMF_i(t) + r(t) \tag{2}$$

where $x(t)$ is the original signal, N is the number of $IMFs$, and r is the residue.

In this study, historical electricity price signal, natural gas price signal and electricity load forecasts were used as input features for the forecasting model. The electricity price signal fluctuations depend on various factors such as weather conditions, electrical grid limitations, and complex market operations. Due to its high nonlinearity, volatility, and nonstationarity, decomposing this signal can enhance the performance of the forecasting model. Hence, in this study, the EMD algorithm was proposed to extract deep features from the price signal, resulting in six IMFs and a residue signal. Another uncertain input of the model was natural gas prices. EMD was also employed to decompose the natural gas price dataset into IMFs and a residue signal.

2.3. CatBoost

The powerful and versatile gradient boosting-based ML algorithm called CatBoost was introduced by [30] and has become famous for its ability to handle categorical features in datasets. Apart from handling categorical inputs, CatBoost offers several benefits compared to other machine learning algorithms. Firstly, the input features are automatically scaled, which can enhance the model's performance. Secondly, CatBoost uses randomized permutations to mitigate the impact of individual variables on the model and prevent overfitting. Lastly, the algorithm can automatically detect and handle missing values in the dataset. With its precision and speed, CatBoost is a powerful and flexible ML algorithm that outperforms alternative algorithms.

The general overview of the Catboost algorithm can be summarized with the following steps:

Step 1- Data Permutation: Algorithm starts with a training dataset 'D' containing 'n' instances. To create diversity, it randomly shuffles the D, d times, generating d different training sets D_r (where r ranges from 1 to d).

Step 2- Matrix Initialization: Algorithm sets up a matrix M where each element $M(r,i)$ represents the initial prediction value for an instance i in training set D_r. Algorithm initializes these values to zero.

Step 3- CatBoost Training on a Random Set: Algorithm randomly chooses one of the permutation sets, D_r, for the following steps.

- Categorical Feature Encoding: Algorithm enhances categorical features using Ordered Target Statistics (TS) Encoding.

- Tree Construction: Algorithm builds a new Ordered Boosting tree (T). This tree approximates the gradient or residual of each instance in D_r utilizing the $M(r,)$ matrix during gradient calculations.

- Gradient Boosting Update: Algorithm uses the newly created tree T to predict outcomes for all permutation datasets. Then updates M based on these predictions using a gradient boosting strategy.

Step 4- Ensemble Prediction: Algorithm repeats the entire Step 3 process N times to build N trees. Finally, algorithm makes predictions for any instance by averaging the predictions from all N trees. This ensemble approach is similar to traditional Gradient Boosting.

In this study, CatBoost was proposed as the regression algorithm for the developed day-ahead electricity price forecasting model. To demonstrate its efficiency on electricity price forecasting, the results of CatBoost were compared with several benchmark algorithms such as linear regression (LR), rifge regression (Ridge), gradient boosting (GB), and SVM.

2.3. Proposed Forecasting Model

This subsection outlines the general architecture of the proposed forecasting model, which consists of three main input features, as illustrated in Figure 1. The general model structure comprises two main stages. In the first stage, the EMD algorithm was used to extract new features from the historical electricity price data and the natural gas data, while the electricity load forecast data was transferred to the second stage. To capture temporal dependencies in sequential data, lagged electricity price and natural gas price inputs were incorporated in the second stage. Finally, the CatBoost algorithm was employed to generate electricity price forecasts.

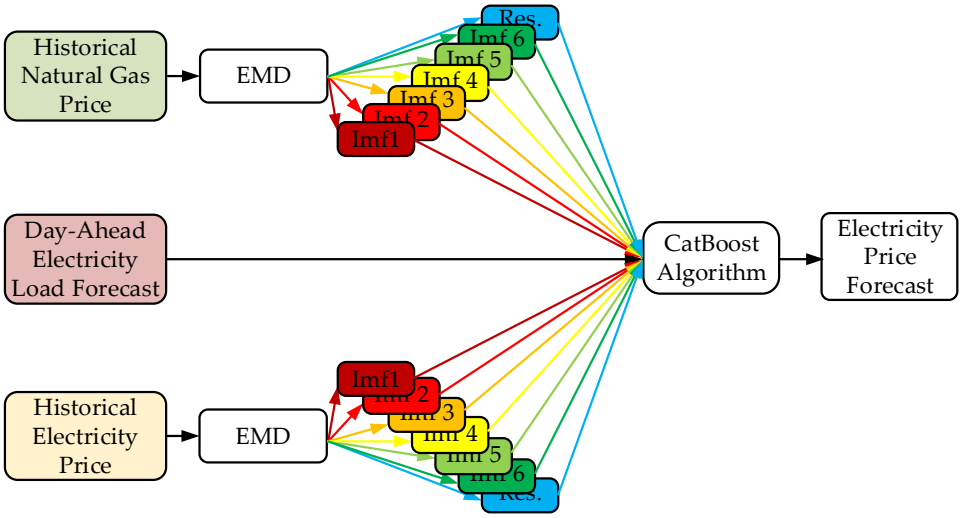


Figure 1. The proposed forecasting model framework

3. DATA DESCRIPTION

In this study, the dataset from the Turkish electricity market was used for developing and testing the proposed forecasting model. The dataset covers hourly samples for a one-year period from October 1, 2019 to December 31, 2019. The total of eight variables—wind, solar, hydro-dam, hydro-river, geothermal generation, day-ahead electricity load forecast, natural gas price, and electricity price values—constituted the dataset. The dataset is publicly available on the online data-sharing platform [31] of the Turkish electricity market. The MI-based feature selection method was used to select important features from the dataset. Figure 2 shows the MI values of the features. Two features with the highest MI values (load forecast and natural gas price) were selected.

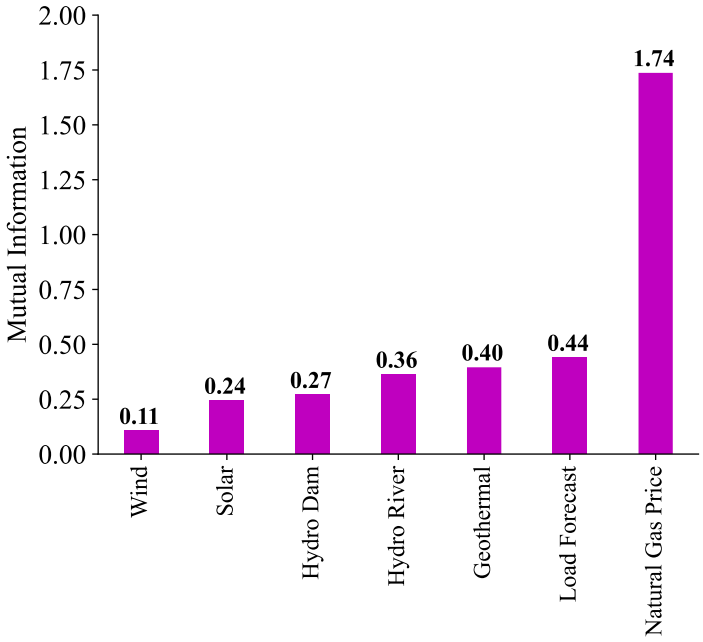


Figure 2. Mutual information between features and electricity price

Figure 3 displays the entire dataset, which consists of the target variable (electricity price (TL)) and the selected model input features (electricity load forecast (MW) and natural gas price (TL)). To provide an in-depth understanding of the dataset, some crucial statistical measures are presented in Table 1.

Table 1. The statistical properties of the dataset

Property	Natural Gas Price	Electricity Load Forecast	Electricity Price
Count	8760	8760	8760
Mean	1479.86	33161.91	260.32
Std.	47.03	4706.18	83.99
Min.	1399.18	18000.00	0.00
25%	1448.30	29300.00	229.99
50%	1474.86	33300.00	300.78
75%	1504.51	36700.00	313.98
Max.	1675.01	45100.00	500.00

The statistics table summarizes the statistical properties of the natural gas price, electricity load forecast, and electricity price datasets. The 8760 samples that made up the dataset included one entire year of hourly data collection. The Count row shows that there were no missing values in any of the three variables. The Mean row shows the average value of each variable. For example, the mean natural gas price was 1479.86 TL, the mean electricity load forecast was 33161.91 MW, and the mean electricity price was 260.32 TL. The Std. row shows the standard deviation of each variable, which is a measure of the spread of the data around the mean. For example, the standard deviation of the natural gas price was 47.03 TL, the standard deviation of the electricity load forecast was 4706.18 MW, and the standard deviation of the electricity price was 83.99 TL. The Min. and Max. rows show the minimum and maximum values of each variable, respectively. For example, the minimum natural gas price was 1399.18 TL, the minimum electricity load forecast was 18000 MW, and the minimum electricity price was 0 TL. The maximum natural gas price was 1675.01 TL, the maximum electricity load forecast was 45100 MW, and the maximum electricity price was 500 TL. The 25%, 50%, and 75% rows show the values that divide the data into quarters, also known as the first quartile, median, and third quartile, respectively. For example, the first quartile of natural gas prices was 1448.30 TL, the median was 1474.86 TL, and the third quartile was 1504.51 TL.

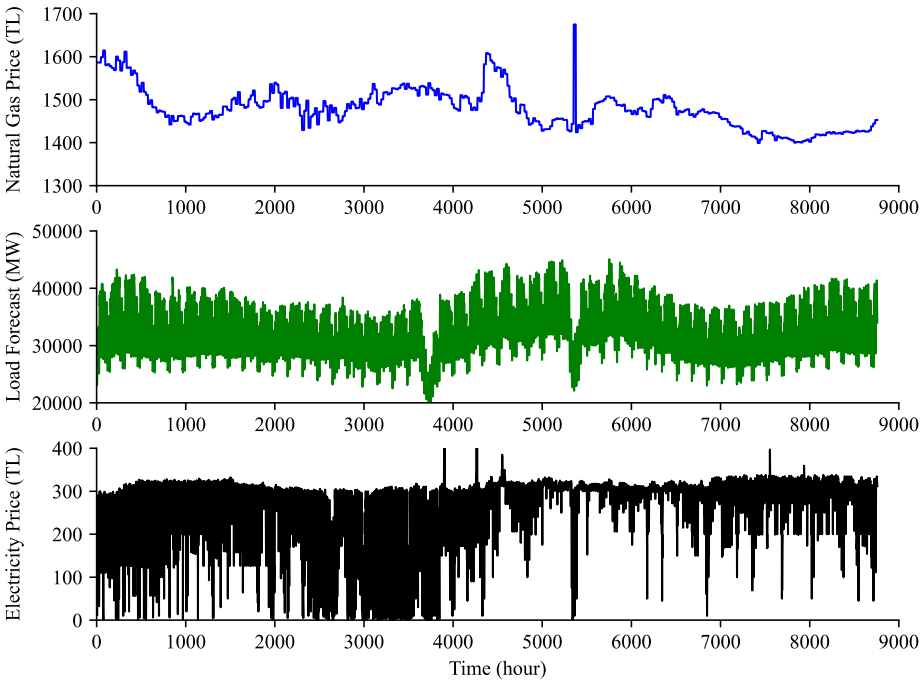


Figure 3. The whole dataset of electricity price, load forecast, and natural gas price

4. EMD BASED FEATURE EXTRACTION

In this study, electricity price data and natural gas price data were decomposed in order to extract more deep features and improve forecasting accuracy. The electricity price data has volatile, nonlinear, and non-stationary characteristics. Several factors, such as complex electricity market operations, weather conditions, and fluctuating renewable generation, have an impact on electricity prices. In addition, natural gas price data has an uncertain nature, as can be seen in Figure 3. Therefore, deep features extracted from these datasets can represent important components of the data. The effective signal decomposition technique called EMD was employed in this study for decomposition. The extracted six IMFs and residue signals are presented in Figure 4. To identify the properties of these signals, some important statistical measures are given in Table 2.

Table 2. The statistical properties of the decomposed signals

Signal	Property	Imf1	Imf2	Imf3	Imf4	Imf5	Imf6	Residue
Electricity Price	Count	8760	8760	8760	8760	8760	8760	8760
	Mean	-0.51	0.67	0.60	0.40	0.43	0.88	257.74
	Std.	26.96	31.08	33.14	33.01	21.15	23.76	52.68
	Min.	-118.41	-147.25	-174.82	-143.19	-75.09	-107.14	93.32
	Max.	115.86	143.20	167.99	143.39	90.93	101.21	335.57
Natural Gas Price	Count	8760	8760	8760	8760	8760	8760	8760
	Mean	0.02	-0.16	-0.11	0.31	-1.17	-3.35	1484.32
	Std.	8.10	12.15	8.19	12.17	13.95	27.72	37.91
	Min.	-110.75	-151.95	-77.92	-80.84	-50.84	-67.45	1408.08
	Max.	107.07	164.72	68.69	68.85	48.70	74.15	1526.15

Table 2 summarizes the statistical properties of the decomposed signals of the electricity price and natural gas price variables using the EMD algorithm. The dataset consists of six intrinsic mode functions (IMFs) and one residue signal for each variable. The statistical properties for each signal are shown, including the number of data points, the average value, standard deviation, minimum, and maximum values. For the electricity price variable, the mean values of IMF1 to IMF6 range from -0.51 to 0.88, and the mean of the residue signal is 257.74. The standard deviation values of these signals range from 21.15 to 33.14, with the residue signal having the highest std. of 52.68. The minimum and maximum values of the IMFs and the residue signal range from -174.82 to 115.86 and 93.32 to 335.57, respectively. For the natural gas price variable, the mean values of IMF1 to IMF6 range from -3.35 to 0.31, and the mean of the residue signal is 1484.32. The standard deviation values of these signals range from 8.10 to 27.72, with the residue signal having the highest std. of 37.91. The minimum and maximum values of the IMFs and the residue signal range from -151.95 to 164.72 and 1408.08 to 1526.15, respectively. Overall, Table 2 provides a summary of the statistical properties of the decomposed signals, which can be useful for understanding the characteristics and behavior of the electricity and natural gas price variables.

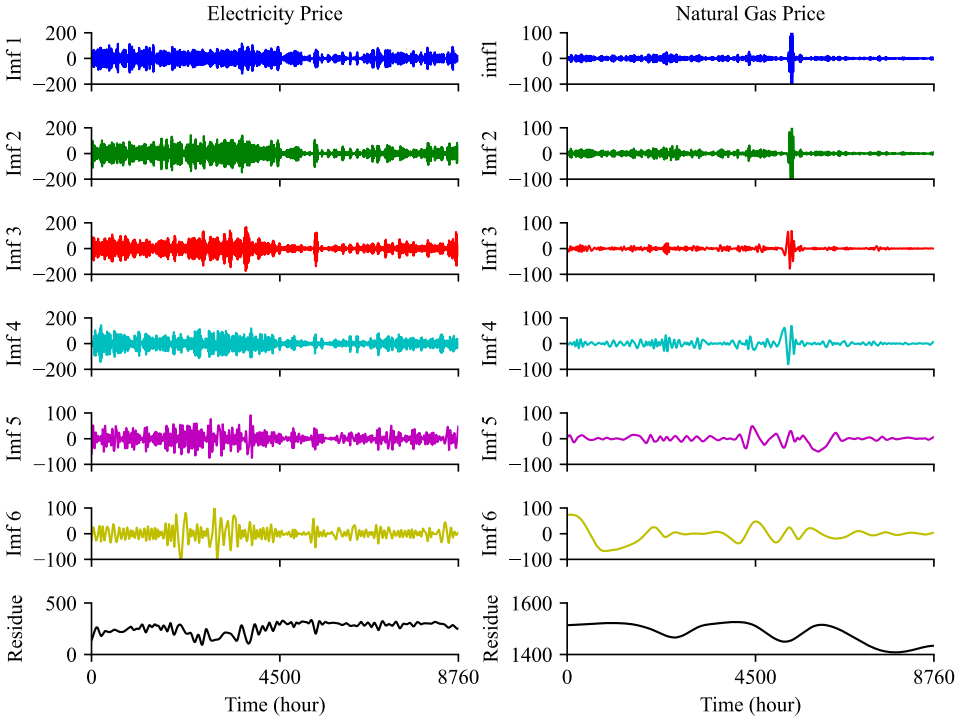


Figure 4. EMD results for electricity price and natural gas price datasets

5. EXPERIMENTAL STUDY

In this study, a two-stage hybrid model was proposed for day-ahead electricity power forecasting. The model was developed and tested on a dataset from the Turkish electricity market. This dataset included four types of features: renewable power generation, day-ahead electricity load forecasts, electricity prices, and natural gas prices. To reduce the number of features, the MI-based feature selection method was used. The two exogenous features, electricity load forecasts, and natural gas prices were selected from these features. Historical electricity prices, historical natural gas prices, and day-ahead electricity load forecasts were used as model input features. The EMD algorithm was employed to extract more deep features from the electricity price signal and the natural gas price signal. The lagged deep electricity price components, the lagged natural gas price components, and electricity load forecasts were used as the regression model inputs. Table 3 presents the model inputs.

Table 3. Inputs for regression

Input		Description	Number of features
Electricity price	$Imf_{1\text{-}6}$	24, 48, 168 hours prior to the target hour	18
	Residue	24, 48, 168 hours prior to the target hour	3
Natural gas price	$Imf_{1\text{-}6}$	24, 48 hours prior to the target hour	12
	Residue	24, 48 hours prior to the target hour	2
Electricity load forecasts		Forecast for target hour	1

The min-max normalization technique was used to scale the model inputs and output values to a range between 0 and 1. 90% of the dataset was allocated for developing (training and validating) the proposed model, whilst the remaining 10% was used for testing. The computationally efficient and effective CatBoost algorithm was used for regression. The Python programming language and Jupyter notebook environment were used to develop forecasting models. The open-source tools NumPy, Pandas, and Scikit-Learn were utilized to implement data preprocessing and ML algorithms. The experimental studies were conducted using the Google Colab platform. The benchmark models based on LR, Ridge, GB, and SVM were compared with the proposed model. The default hyper parameters provided by the

Scikit-Learn library were used in the study. Three error-based metrics, RMSE, MAE, and MAPE, and the correlation coefficient R were used to evaluate performances. The following equations were used to calculate performance metrics.

$$\mathrm{RMSE}(\%) = 100 \times \sqrt{\frac{1}{N}\sum_{h=1}^{N}\left(P_a^h - P_f^h\right)} \tag{3}$$

$$\mathrm{MAE}(\%) = 100 \times \frac{1}{N}\sum_{h=1}^{N}\left|P_a^h - P_f^h\right| \tag{4}$$

$$\mathrm{MAPE}(\%) = 100 \times \frac{1}{N}\sum_{h=1}^{N}\left|\frac{P_a^h - P_f^h}{P_a^h}\right| \tag{5}$$

$$\mathrm{R}(\%) = 100 \times \frac{\sum_{h=1}^{N}\left(P_f^h - \overline{P_f}\right)\left(P_a^h - \overline{P_a}\right)}{\sqrt{\sum_{h=1}^{N}\left(P_f^h - \overline{P_f}\right)^2 \sum_{h=1}^{N}\left(P_a^h - \overline{P_a}\right)^2}} \tag{6}$$

where N is the number of samples, a and f are the actual and forecasted values, respectively, and h denotes the hour. $\overline{P_a}$ and $\overline{P_f}$ are the averages of the actual and forecasted price values.

6. RESULTS

In this study, the results of the proposed and benchmark models were evaluated on a real dataset from the Turkish electricity market. The test results for both benchmark and proposed models are illustrated at the Figure 5.a, showing that all the forecasting models converged to the actual price values. To better demonstrate the forecasting performances, small intervals from the test data were selected and presented at the Figure 5.b and c. These two graphics demonstrated that the proposed CatBoost algorithm provided improved accuracy.

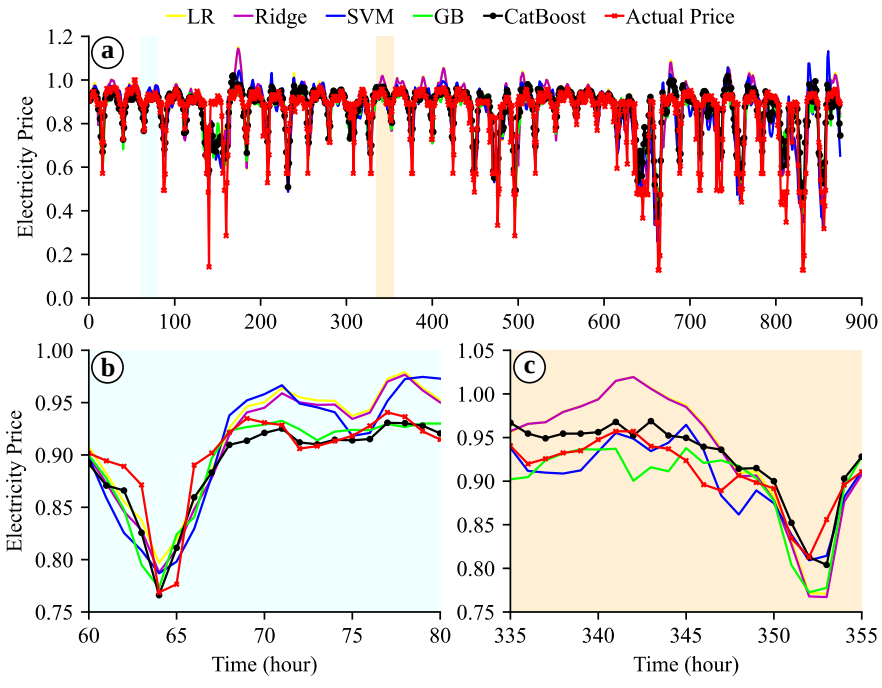


Figure 5. a. Overall results, b. Results for specific time interval, c. Results for specific time interval

Table 4 presents the forecasting results of benchmark and proposed models. Four performance metrics were calculated to evaluate the performance of the models. RMSE% is a well known error based metric. A lower RMSE% value indicates better accuracy in the model's predictions. In Table 4, the method with the lowest RMSE is CatBoost (8.3282%), indicating that it had the most accurate predictions among the methods tested. MAE represents the mean absolute difference between the actual values and the forecasted values. A lower MAE value indicates better accuracy in the model's predictions. In Table 4, the method with the lowest MAE is also CatBoost (5.2210%), indicating that it had the most accurate predictions among the methods tested. MAPE represents the mean absolute percentage difference error. A lower MAPE value indicates better accuracy in the model's predictions. In Table 4, the method with the lowest MAPE is once again CatBoost (6.9675%), indicating that it had the most accurate predictions among the methods tested. R% represents the coefficient of determination for actual and forecasted values. A higher R% value indicates that the model is better at predicting the actual values. In Table 4, the method with the highest R% value is CatBoost (86.2256%), indicating that it had the best predictive performance among the methods tested. Overall, the table illustrates that CatBoost outperformed the other methods in terms of accuracy and predictive power, as evidenced by its consistently lower RMSE, MAE, and MAPE values and higher R value.

Table 4. Forecasting results

Methods	Metrics			
	RMSE%	MAE%	MAPE%	R%
LR	9.0803	6.5729	8.7026	81.6213
Ridge	8.9713	6.4423	8.5794	81.8418
SVM	8.9384	6.1627	8.4436	81.5965
GB	8.5352	5.5233	7.4667	84.6193
CatBoost	8.3282	5.2210	6.9675	86.2256

To demonstrate the impact of EMD on forecasting performance, a further analysis was carried out on the results. The improvement in performance was calculated by comparing the benchmark and proposed models with the LR model without EMD. The results of this improvement in performance are presented in a bar graphic given in Figure 6.

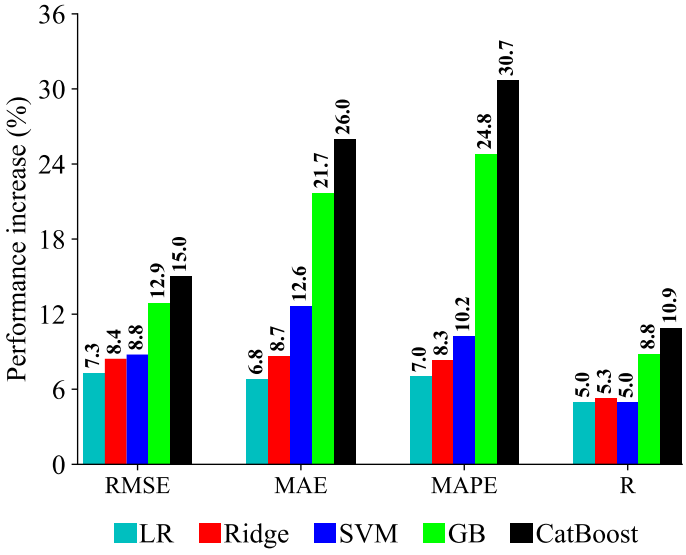


Figure 6. Performance incensements compared to the LR without EMD

The bar graphic shows the performance increases in four metrics (RMSE, MAE, MAPE, and R) for five different methods (LR, Ridge, GB, SVM, and CatBoost) compared to the LR model without EMD decomposition. The performance increases are expressed as a percentage increase in each metric.

According to bar graphic given, it can be interpret that the LR and Ridge had the smallest performance increase among all methods, with less than 10% increase in all metrics. The SVM and GB models showed moderate performance improvements, with 8.8-12.9% increase in RMSE, 12.6-21.7% increase in MAE, 10.2-24.8% increase in MAPE, and 5.0-8.8% increase in R. The proposed CatBoost method showed the largest improvement in all metrics, with a 15.8% increase in RMSE, 26.0% increase in MAE, 30.7% increase in MAPE, and 10.9% increase in R. This suggests that the CatBoost method outperformed the other methods and had the best performance improvement over the LR model, based on the given metrics.

7. CONCLUSION

In this study a two stage hybrid model was proposed for day ahead electricity price forecasting. MI and EMD were used in the first stage for feature selection and feature extraction, respectively. Day ahead electricity load forecasts, time lagged natural gas prices and time lagged electricity price signal components were used as second stage (regression model) inputs. The CatBoost algorithm was proposed for the second stage. LR, Ridge, GB, and SVM were used as benchmark regression algorithms. The implementation of the forecasting models were realized on the Turkish electricity market data. The experimental studies showed proposed forecasting model based on Catboost algorithm outperformed the benchmark models. The proposed model achieved minimum RMSE, MAE, and MAPE as 8.3282%, 5.2210%, and 6.9675% and maximum R value as 86.2256%. In addition the performance increase analysis confirmed the using EMD based decomposition improves the forecasting accuracy. The results of this study indicated that the proposed two-stage hybrid model has promising potential in forecasting day-ahead electricity prices in the Turkish electricity market. The future work that considers incorporating additional data sources, such as weather data or social media data, to further enhance the forecasting model will be interesting.

Declaration of Ethical Standards

The author of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Credit Authorship Contribution Statement

Ceyhun Yıldız, conceptualized, conducted, and wrote the entire study.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding / Acknowledgements

This study was not funded by any institution.

Data Availability

The author confirms that the datasets analyzed during the current study are available in public repositories (see reference [31]).

7. REFERENCES

[1] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark," *Appl. Energy*, vol. 293, no. December 2020, p. 116983, 2021, doi: 10.1016/j.apenergy.2021.116983.

[2] R. Weron, "Electricity price forecasting : A review of the state-of-the-art with a look into the future," *Int. J. Forecast.*, vol. 30, no. 4, pp. 1030–1081, 2014, doi: 10.1016/j.ijforecast.2014.08.008.

[3] L. Tschora, E. Pierre, M. Plantevit, and C. Robardet, "Electricity price forecasting on the day-ahead market using machine learning," *Appl. Energy*, vol. 313, no. March, p. 118752, 2022, doi: 10.1016/j.apenergy.2022.118752.

[4] P. Wang *et al.*, "An Online Electricity Market Price Forecasting Method Via Random Forest," *IEEE Trans. Ind. Appl.*, vol. 58, no. 6, pp. 7013–7021, 2022.

[5] C. Xiao, D. Sutanto, K. M. Muttaqi, M. Zhang, K. Meng, and Z. Y. Dong, "Online sequential extreme learning machine algorithm for better predispatch electricity price forecasting grids," *IEEE Trans. Ind. Appl.*, vol. 57, no. 2, pp. 1860–1871, 2021.

[6] I. Shah, H. Bibi, S. Ali, L. Wang, and Z. Yue, "Forecasting one-day-ahead electricity prices for italian electricity market using parametric and nonparametric approaches," *IEEE Access*, vol. 8, pp. 123104–123113, 2020.

[7] H. Yang and K. R. Schell, "GHTnet : Tri-Branch deep learning network for real-time electricity price forecasting," *Energy*, vol. 238, p. 122052, 2022, doi: 10.1016/j.energy.2021.122052.

[8] H. Yang and K. R. Schell, "International Journal of Electrical Power and Energy Systems QCAE : A quadruple branch CNN autoencoder for real-time electricity price forecasting," *Int. J. Electr. Power Energy Syst.*, vol. 141, no. April, p. 108092, 2022, doi: 10.1016/j.ijepes.2022.108092.

[9] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Appl. Energy*, vol. 221, no. April, pp. 386–405, 2018, doi: 10.1016/j.apenergy.2018.02.069.

[10] S. Luo and Y. Weng, "A two-stage supervised learning approach for electricity price forecasting by leveraging different data sources," *Appl. Energy*, vol. 242, no. February, pp. 1497–1512, 2019, doi: 10.1016/j.apenergy.2019.03.129.

[11] T. Zhang, Z. Tang, J. Wu, X. Du, and K. Chen, "Short term electricity price forecasting using a new hybrid model based on two-layer decomposition technique and ensemble learning," *Electr. Power Syst. Res.*, vol. 205, no. July 2021, p. 107762, 2022, doi: 10.1016/j.epsr.2021.107762.

[12] A. Meng *et al.*, "Electricity price forecasting with high penetration of renewable energy using attention-based LSTM network trained by crisscross optimization," *Energy*, vol. 254, p. 124212, 2022, doi: 10.1016/j.energy.2022.124212.

[13] G. Memarzadeh and F. Keynia, "Short-term electricity load and price forecasting by a new optimal LSTM-NN based prediction algorithm," *Electr. Power Syst. Res.*, vol. 192, no. November 2020, p. 106995, 2021, doi: 10.1016/j.epsr.2020.106995.

[14] Z. Shao, Q. Zheng, C. Liu, S. Gao, G. Wang, and Y. Chu, "A feature extraction- and ranking-based framework for electricity spot price forecasting using a hybrid deep neural network," *Electr. Power Syst. Res.*, vol. 200, no. September 2020, p. 107453, 2021, doi: 10.1016/j.epsr.2021.107453.

[15] X. Xiong and G. Qing, "A hybrid day-ahead electricity price forecasting framework based on time series," *Energy*, vol. 264, no. November 2022, p. 126099, 2023, doi: 10.1016/j.energy.2022.126099.

[16] K. Bhatia, R. Mittal, J. Varanasi, and M. M. Tripathi, "An ensemble approach for electricity price forecasting in markets with renewable energy resources," *Util. Policy*, vol. 70, no. July 2020, p. 101185, 2021, doi: 10.1016/j.jup.2021.101185.

[17] M. Heidarpanah, F. Hooshyaripor, and M. Fazeli, "Daily electricity price forecasting using artificial intelligence models in the Iranian electricity market," *Energy*, vol. 263, no. PE, p. 126011, 2023, doi: 10.1016/j.energy.2022.126011.

[18] S. Demir, K. Mincev, K. Kok, and N. G. Paterakis, "Data augmentation for time series regression : Applying transformations , autoencoders and adversarial networks to electricity price forecasting

☆," *Appl. Energy*, vol. 304, no. September, p. 117695, 2021, doi: 10.1016/j.apenergy.2021.117695.

[19] W. Qiao and Z. Yang, "Forecast the electricity price of U . S . using a wavelet transform-based hybrid model," *Energy*, vol. 193, p. 116704, 2020, doi: 10.1016/j.energy.2019.116704.

[20] K. Iwabuchi, K. Kato, D. Watari, I. Taniguchi, and F. Catthoor, "Energy and AI Flexible electricity price forecasting by switching mother wavelets based on wavelet transform and Long Short-Term Memory," *Energy AI*, vol. 10, no. May, p. 100192, 2022, doi: 10.1016/j.egyai.2022.100192.

[21] D. H. Vu, K. M. Muttaqi, A. P. Agalgaonkar, and A. Bouzerdoum, "Short-term forecasting of electricity spot prices containing random spikes using a time-varying autoregressive model combined with kernel regression," *IEEE Trans. Ind. Informatics*, vol. 15, no. 9, pp. 5378–5388, 2019.

[22] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. A. Kaboli, and S. Ahmad, "Short-Term Electricity Price Forecasting via Hybrid Backtracking Search Algorithm and ANFIS Approach," *IEEE Access*, vol. 7, pp. 77674–77691, 2019, doi: 10.1109/ACCESS.2019.2922420.

[23] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. R. A. Kaboli, S. Ahmad, and S. P. Ang, "Hybrid ANN and artificial cooperative search algorithm to forecast short-term electricity price in de-regulated electricity market," *Ieee Access*, vol. 7, pp. 125369–125386, 2019.

[24] N. Bibi, I. Shah, A. Alsubie, S. Ali, and S. A. Lone, "Electricity Spot Prices Forecasting Based on Ensemble Learning," *IEEE Access*, vol. 9, pp. 150984–150992, 2021, doi: 10.1109/ACCESS.2021.3126545.

[25] A. L. I. N. Alkawaz and A. Abdellatif, "Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model," *IEEE Access*, vol. 10, no. October, pp. 108021–108033, 2022, doi: 10.1109/ACCESS.2022.3213081.

[26] S. Zhou, L. Zhou, M. Mao, H.-M. Tai, and Y. Wan, "An optimized heterogeneous structure LSTM network for electricity price forecasting," *IEEE Access*, vol. 7, pp. 108161–108173, 2019.

[27] R. Zhang, G. Li, and Z. Ma, "A deep learning based hybrid framework for day-ahead electricity price forecasting," *IEEE Access*, vol. 8, pp. 143423–143436, 2020.

[28] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[29] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[31] "Energy Exchange Istanbul," 2023. https://www.epias.com.tr.