A Novel Hybrid Model for Electricity Price Forecasting Based on the Integration of Bi-Directional Long Short-Term Memory and Gated Recurrent Unit

Amirhossein Hajigholam Saryazdi

A Thesis

in

the Department

of

John Molson School of Business

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Business Analytics and Technology Management at

Concordia University

Montreal, Quebec, Canada

October 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Amirhossein Hajigholam Saryazdi

Entitled: A Novel Hybrid Model for Electricity Price Forecasting Based on the Integration of Bi-Directional Long Short-Term Memory and Gated Recurrent Unit

and submitted in partial fulfillment of the requirements for the degree of

Master of Science in Business Analytics and Technology Management

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. Arman Sadreddin	Chair
Dr. Dongliang Sheng	Examiner
Dr. Rustam Vahidov	Examiner
Dr. Salim Lahmiri	Supervisor

Approved by: Dr. Suchit Ahuja, Graduate Program Director

1st November , 2024 Dr. Anne-Marie Croteau, Dean of Faculty

Abstract

A Novel Hybrid Model for Electricity Price Forecasting Based on the Integration of Bi-Directional Long Short-Term Memory and Gated Recurrent Unit

Amirhossein Hajigholam Saryazdi

The prediction of electricity prices plays a pivotal role in the wholesale electricity markets, influencing sale prices, bidding strategies, electricity dispatch, control, and the management of market. Notably, forecasting in a deregulated electricity market is challenging due to multiple factors such as high volatility, non-stationarity and multi-seasonality of electricity prices. In response to this challenge, this research proposes a novel hybrid deep learning model employing Bi-directional Long Short-Term Memory (Bi_LSTM) and Gated Recurrent Unit (GRU) for real-time electricity price forecasting. In this model, the output sequences from the Bi_LSTM layer, which captures both past and future temporal dependencies, are directly fed into the GRU layer to refine the feature extraction. This hybrid approach not only reduces overfitting risk of a single model, but also increases robustness and adaptability of model. Three studies are conducted in New York City (NYC), electricity market to evaluate the model by systematically comparing the obtained results. First, the proposed model, Bi_LSTM-GRU, outperforms several baseline models, spanning a statistical time-series method: Auto Regressive Integrated Moving Average (ARIMA), Machine Learning approaches: Linear Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Support Vector Regression (SVR), and Deep Learning techniques: Long Short-Term Memory (LSTM), Bi-LSTM, GRU, and Convolutional Neural Network (CNN). Secondly, the possibility of hybridizing CNN and Recurrent Neural Network (RNN) architectures has been examined. The proposed model also surpasses CNN-LSTM, CNN-Bi-LSTM, and CNN-GRU. Lastly, the potential contribution of data decomposition techniques in enhancing the proposed model has been assessed. It is found out that adding Wavelet Transform (WT) or Fourrier Transform (FT) to decompose the data leads to higher error rates.

***Keywords**:* Electricity Price Forecasting, Deep Learning, Comparative study, Interconnected Grids, Power Market

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Dr. Salim Lahmiri, for his invaluable guidance, encouragement, and support throughout this research. Your insightful feedback and expertise were essential in helping me refine my ideas and shape this thesis.

A special thanks to Dr. Arman Sadreddin, the chair of my committee, for ensuring the smooth progress of my thesis evaluation. Your assistance throughout the process is much appreciated.

I am also grateful to my examiner, Dr. Dongliang Sheng, for accepting the role. Your feedback is essential to improve the quality of this work.

I would also like to extend my sincere thanks to Dr. Rustam Vahidov, Chair of the Department of Supply Chain and Business Technology Management, for taking the time to serve as my examiner. Your expert feedback and valuable insights are instrumental in further enhancing the quality of my work.

This thesis would not have been possible without the support and guidance of each of you. Thank you!

Table of Contents

List of Figures

List of Tables

Acronyms

EPF	Electricity Price Forecasting
NYC	New York City
ARIMA	Auto-regressive Integrated Moving Average
LR	Linear Regression
RF	Random Forest
SVR	Support Vector Regression
XGB	Extreme Gradient Boosting
LSTM	Long Short-Term Memory
Bi-LSTM	Bi-directional Long Short-Term Memory
AM	Attention Mechanism
GRU	Gated Recurrent Unit
CNN	Convolutional Neural Networks
ML	Machine Learning
DL	Deep Learning
DE	Differential evolution
SAE	Stacked autoencoder
CSO	crisscross optimization algorithm
VMD	Variational Mode Decomposition
FEEMD	Fast ensemble empirical mode decomposition
DAP	Day-ahead Price
RTP	Real-time Price
NP	Nord Pool
PJM	Pennsylvania–New Jersey-Maryland
ACBFS	Adaptive copula based mutual information
RNN	Recurrent Neural Networks
Val	Validation
FFT	Fast Fourier Transform
WT	Wavelet Transform

1 Introduction

As the power industry undergoes liberalization and deregulation, trading activities within the power market exhibit a growing array of concealed dynamics and uncertainties. In this situation, electricity price has is the dominant player in the market, and accurate electricity price forecasting (EPF) has become essential for market participants. Firstly, it can help power generation companies to obtain higher economic profit by selecting bidding strategies in day-ahead market. Secondly, customers can have a better electricity consumption plan according to the predicted price. Finally, having a reference of electricity price prediction, can not only enable market regulators to prevent other players with high market power from manipulating electricity price but also help system operators to ensure electric grid reliability (Jiang et al., 2023) (Meng et al., 2022).

In comparison to other commodities such as crude oil and natural gas, electricity exhibits certain different characteristics. To name a few, the electricity demand is relatively inelastic; it cannot be stored in large quantities at reasonable price, so a constant balance between electricity generational and consumption is crucial to prevent blackouts or overloading the grid; and electricity supply is a combination of flexible, inflexible, intermittent, and volatile plants, thus complicating the energy distribution (Lehna et al., 2022). Considering these factors, electricity prices demonstrate intrinsic traits that pose challenges in forecasting. These include high frequency and volatility, unstable mean and variance, multi-seasonality, nonlinearity, negative prices and extreme highs, referred to as spikes, a phenomenon uncommon in other commodity markets (Ehsani et al., 2024).

Interconnected grids enable operators to have electricity exchange, known as import/export, with neighboring markets. Contrary to Europe where an operator runs the whole power market in country (e.g. Spain) or even a group of countries (e.g. Nordpool), there are multiple market operators in both the US and Canada. Among those in the US, New York System Operator (NYISO) is the only market that has import and export through interconnections with two Canadian markets, namely Ontario and Quebec, and two US markets namely New England and

Pennsylvania–New Jersey-Maryland (PJM) (FERC, 2024). This unique characteristic qualifies NYISO as a challenging dynamic market for being explored in terms of EPF.

This study proposes employing Bi-LSTM-GRU model, a hybrid model based on RNN architecture, in the New York electricity market. The proposed model is evaluated and compared to commonly used Machine Learning (ML) and Deep Learning (DL) and statistical models using prediction performance metrics. To the best of the author's knowledge, this research is one of the initial attempts to apply and evaluate this method in the electricity market.

1.1 Background

The electricity market generally comprises generators, transmission operators, distributors, and end users. It can be regulated, where a single entity owns and operates all these components and sets rates (e.g., Hydro-Québec in Quebec). Alternatively, in deregulated markets, Independent System Operators (ISOs) manage the transmission grid and coordinate market operations (e.g., IESO in Ontario, NYISO in New York). ISOs play a vital role in overseeing the electricity market, fostering competition, and planning to meet future demand with secure resources. One of their primary responsibilities is ensuring grid reliability and balancing supply and demand by receiving and evaluating bids and offers from generators and consumers. Generators submit offers indicating the price at which they are willing to produce electricity, while consumers (or load-serving entities) submit bids reflecting their demand. Figure 1 indicates the position of system operator in a deregulated power market (Hajigholam Saryazdi, 2024a)

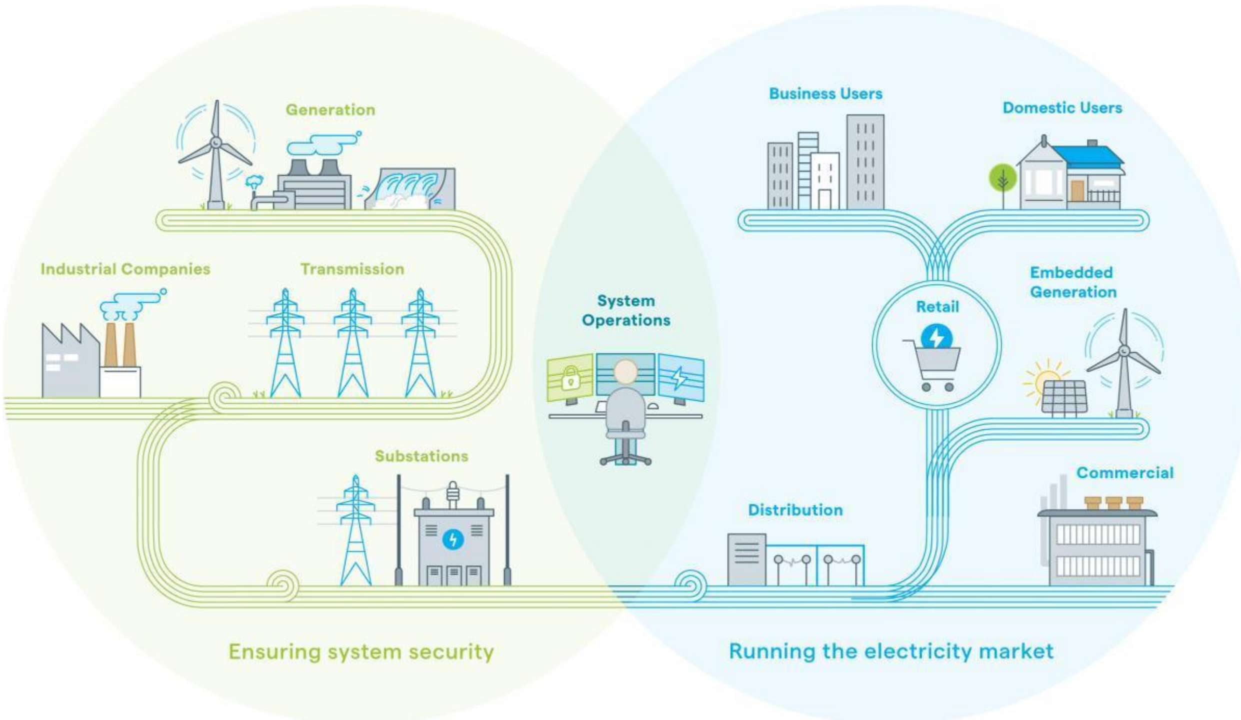

Figure 1: The position of system operator in a deregulated power market (transpower.co.nz, 2024)

Wholesale trading involves bulk buying and selling of electricity through several market mechanisms. The Day-Ahead Market (DAM) allows participants to commit to supply or consumption a day before delivery, ensuring they avoid being negatively affected by market volatility. The Real-Time Market (RTM) addresses immediate balancing needs, letting market participants sell and buy wholesale electricity throughout the course of an operating day with no set or previously agreed prices but a price that reflects market demand. Bilateral contracts involve private agreements between entities, where both the buyer and seller negotiate specific terms such as the price, quantity, and delivery schedule of electricity, providing flexibility and often used for long-term commitments or tailored supply arrangements. Accurate year-ahead price forecasting enables market participants to better anticipate future price movements, allowing for more informed decisions, particularly in bilateral contracts and market trades, ultimately enhancing strategic planning and financial stability across the power market (Hajigholam Saryazdi, 2024b).

1.2 Motivation and Contribution

The contribution of this study is four-fold in terms of selecting the forecasting horizon, market, variables, availability, and model:

I. Although numerous models have been proposed for EPF, they are limited to the market and timespan of their study due to the unique characteristics of each market and different pricing mechanism (Lago et al., 2021). Hence, each market requires separate and up-to-date research. New York electricity market, which is the scope of this paper, shows lack of price forecasting literature comparing to other electricity markets worldwide (Aggarwal et al., 2009).

II. Interconnected electricity markets are one of the key factors that influence the electricity price (Lago, De Ridder, Vrancx, et al., 2018) yet among the least explored in the literature (Lu et al., 2021). This paper aims to bridge this gap by selecting a market that is interconnected with four other markets.

III. Many previous researchers have not shared their data or model preventing academic community and practitioners to build upon their work (Lago, De Ridder, Vrancx, et al., 2018). However, this study aims to use publicly available data as well as to provide an open-source model for the benefit of future researchers.

IV. Although the practice of using hybrid models have been well-established in the domain of EPF, Bi-LSTM-GRU hybridization has not been previously used for this purpose.

1.3 Literature review

According to the literature (Lu et al., 2021), the prediction model of electricity price can be divided into three modules: data preprocessing method, optimizer , and prediction model. Data preprocessing techniques encompass various methods, including the Wavelet Transform (WT) (Antonini et al., 1992), Variational Mode Decomposition (VMD) (Dragomiretskiy & Zosso, 2014), Empirical Mode Decomposition (EMD) (N. E. Huang et al., 1998), and their respective variants. These techniques serve as tools for time series decomposition, with the aim of improving prediction accuracy. The objective of an optimizer lies in the refinement of hyperparameters within a prediction model, achieved through the minimization of errors.

Noteworthy among these optimization techniques are metaheuristic algorithms, such as Differential Evolution (DE) (Storn, 1997), recognized for its stochastic search methodology. Additionally, Bayesian Optimization (BO) stands out as a probabilistic model, selecting hyperparameters by estimating the posterior distribution of the objective function through the application of Bayes' theorem (Cheng et al., 2019). The data-driven prediction model is responsible for forecasting and is broadly categorized as statistical, ML/DL, and hybrid models.

Traditional statistical such as Regression Model, Exponential Smoothing (ES), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and vector autoregressive model (VAR), are used for EP prediction. These simple methods consider temporal relationship of data, rendering them suitable for predicting stable series characterized by minor fluctuations and low-frequency changes. Nevertheless, it is noteworthy that most statistical models operate as linear forecasters and may exhibit suboptimal performance when confronted with high-frequency data, such as hourly electricity price fluctuations (Lago, De Ridder, & De Schutter, 2018). The complexity of electricity price dynamics, encompassing complex features like high and low frequencies, volatility, variable means and variances, as well as a substantial occurrence of atypical prices, pose a challenge for conventional methods, constraining their precision in predicting electricity prices (AL-Musaylh et al., 2018).

ML and DL models exhibit proficiency in discerning complex nonlinear features within electricity price datasets. These approaches alleviate the necessity of adhering to statistical assumptions and have demonstrated superior accuracy in forecasting nonlinear time series data. A compilation of models recently employed for EPF is provided in Table 1.

DL models that excel in EPF include Long Short-Term Memory (LSTM) (Cheng et al., 2019; Peng et al., 2018; Xiong & Qing, 2023), Gated Recurrent Unit (GRU) (Lago, De Ridder, & De Schutter, 2018), Attention Mechanism (Meng et al., 2022), Auto-Encoder (AE) (Qiao & Yang, 2020) and Convolution Neural Network (CNN) (Ehsani et al., 2024; Ghimire et al., 2024). Lago, De Ridder, & De Schutter (2018) conducted a comparative analysis of various DL models, including DNN, GRU, and LSTM, along with ML models such as RF and RBFNN, as well as the ARIMA statistical model. They utilized European power exchange (EPEX) Belgium datasets to predict spot EP. The study revealed that DL models outperformed ML models, as indicated by the Symmetric Mean Absolute Percentage Error.

Leveraging optimization techniques, Peng et al. (2018) used DE to fine-tune hyperparameters for LSTM. Through an extensive analysis of electricity markets in New South Wales (NSW), Germany/Austria, and France, conducted for both one-step-ahead and multi-step-ahead forecasting, their DE-LSTM model demonstrated superior performance when compared to ARIMA, ANN, RNN, SVM, and DE_BPNN models, as evidenced by lower MAE, MSE, MAPE, and RMSE.

Harnessing decomposition techniques, Cheng et al. (2019) incorporated WT in conjunction with LSTM and evaluated its performance using datasets from New South Wales, Australia, and France. WT was applied to decompose the data, breaking the EP time-series into several component series with minor variances. These decomposed time-series were individually trained and predicted using LSTM, and the resulting predictions were aggregated to obtain the final forecast. The utilization of WT contributed to the stabilization of the variance in the time-series data, enabling LSTM to more accurately capture fluctuations in EP. This approach significantly improves prediction accuracy compared to a model that combined ARIMA and Artificial Neural Network (ANN) models. Additionally, Cheng et al. highlighted the advantages of the Adam optimizer when paired with LSTM, demonstrating its superiority over Stochastic Gradient Descent (SGD) and RMSProp optimizers in their study.

The methodology for incorporating preprocessing and optimization techniques into prediction models is an evolving domain. Cheng et al. (2019) introduced a progressive hybrid model denoted as EWT-SVR-BiLSTM-BO. In this model, the low-frequency output of Empirical Wavelet Transform (EWT) was directed to a Support Vector Regression (SVR) model, while the high-frequency output was fed into a Bidirectional Long Short-Term Memory (BiLSTM) model. Notably, hyperparameters for the entire ensemble were finely tuned using Bayesian Optimization (BO). This innovative approach resulted in superior performance, as evidenced by lower prediction errors when compared to SVR, Gradient Boosted Decision Trees (GBDT), Extreme Learning Machine (ELM), and BiLSTM models.

Despite the success of individual prediction models, the intricate nature of EP poses challenges in achieving optimal predictions through a single model. Consequently, the adoption of hybrid models has become a prevalent approach for EPF. These hybrid models integrate various ML and DL techniques, leveraging their complementary strengths to attain superior accuracy in prediction. Scholars have introduced diverse hybrid models that exhibit superior performance compared to their individual model counterparts. For instance, in New York market, GRU hybridized with VMD-CNN exhibited superior performance compared to LSTM, CNN, and VMD-CNN models in short-term EPF (C.-J. Huang et al., 2021). In a similar vein, Qiao & Yang (2020) employed WT, Stacked Autoencoder (SAE), and LSTM models to generate price predictions for United State electricity markets. Despite the superior prediction accuracy exhibited by the SAE-LSTM model, the WT-SAE-LSTM model was considered to hold greater practical value surpassing the predictive performance of both LSTM and BiLSTM.

In the context of Danish electricity market EWT proved to be superior to VMD and Ensemble Empirical Mode Decomposition (EEMD) preprocessing when integrated with LSTM. The incorporation of an Attention Mechanism (AM) further reduced errors in the joint EWT-LSTM model. Through optimization by Cuckoo Search Algorithm (CSO), the EWT-AM-LSTM-CSO model achieved lower error rates, surpassing the performance of EWT-SVM, and GRUs (Meng et al., 2022). In a extensive investigation aimed at identifying the optimal combination of feature selection, decomposition, and optimization techniques in the PJM market, Xiong & Qing, (2023) discovered that the ACBFS-VMD-LSTM-BOHB hybrid model emerged as the most effective among various alternatives, particularly in terms of minimizing errors.

Indeed, it is crucial to note that the superiority of hybrid methods is not universally guaranteed, and there are instances where a single prediction model might perform comparably or even outperform its hybrid counterparts. Similarly, the assumption that DL models consistently outshine statistical methods is challenged by (Lehna et al., 2022).

In the pursuit of identifying a winning model, ensemble models, which combine multiple forecasts from the same model calibrated on different windows, have consistently demonstrated significantly better results than their individual counterparts across various electricity markets, including PJM, France, Belgium, Nord pool, and Germany (Lago et al., 2021). Building upon this research, Tschora et al. (2022) have shown that incorporating the price history of neighboring countries significantly enhances the quality of day-ahead forecasting (Tschora et al., 2022). This highlights the potential benefits of leveraging ensemble methods and considering broader regional factors for more accurate and robust electricity price predictions.

An additional innovative approach that can be integrated into a model involves incorporating an error compensation stage. This enhancement has demonstrated improved performance, particularly in real-time half-hourly price prediction within the Australian electricity market (Ghimire et al., 2024).

In the exploration of additional input variables for electricity market forecasting, the inclusion of factors like wind or solar generation, wind speed, temperature, and predicted load has been studied extensively. However, findings from research in the Ontario electricity market suggest that incorporating exogenous variables with low correlation (less than 0.5) can lead to an increase in prediction errors (Ehsani et al., 2024)

Table 1: Selected literature

Author	Market (Year)	Decomp-osition	Base Model	Opti-mizer	Horizon	Target (Input) Variables	Data Split (Size)
(Peng et al., 2018)	NSW1 (2013) Germany/Austria2 (2012-2015) France3 (2017)	_	LSTM	DE	H D H	HEP	700:20:24 876:87:220 1008:144:288
(Chang et al., 2019)	NSW (2013) France (2017) France (2018)	WT	LSTM	Adam	H H	HEP	(744) (100) (1008)
(Cheng et al., 2019)	European Power Exchange	EWT	SVR-BiLSTM	BO	H	HEP	504:216
(Qiao & Yang, 2020)	EIA (US) 4 (1997-2020)	WT	SAE-LSTM	_	M	HEP	221:50:17
(C.-J. Huang et al., 2021)	New York (2015-2018)	VMD	CNN-GRU	_	H	HEP	
(Lehna et al., 2022)	Germany	_	CNN-LSTM	_	H	HEP, Wind speed, Consumer Load, Avg. solar radiation, Fuel price, CO2 emission price	720

[1] www.aemo.com.au/Electricity
[2] www.epexspot.com
[3] www.epexspot.com
[4] www.eia.gov

(Meng et al., 2022)	Denmark (2018-2019)	EWT	AM-LSTM	CSO	H	HEP, Wind Power Generation, Solar power Generation, Predicted Load	730
(Xiong & Qing, 2023)	PJM (2015-2016, 2017)	VMD (ACBFS)	LSTM	BOHB	H	HEP	1272:168 [each season] 2015-2016:2017
(Ghimire et al., 2024)	Queensland (2014-2022)	VMD	CNN-LSTM	_	H-H Real-time	Historical electricity price	~ 110000: 27000: 4000 [seasonal] (13000) [yearly]
(Ehsani et al., 2024)	Ontario (2021-2022)	_	TriCNN-GRU	_	H	HEP, import/export, demand, weather, generations	7066:1766: 168

2 Methodology

This section describes data, base line predictive models, proposed forecasting models, decomposition techniques, and evaluation criteria employed in this research from a technical point of view.

2.1 Data

This study focuses on predicting electricity price in New York, USA. Two main reasons behind this selection are: Firstly, it fills the lack of research in this market relatively to other markets around the world (Aggarwal et al., 2009). Secondly, this data is publicly available by New York Independent System Operator (NYISO)[5] enabling future researchers to freely replicate this research and build upon this line of literature. This market consists of different zones, each has separate regional price. The focus of this paper is NYC (zone J) due to its largest size in terms of both electricity load and population.

NYISO set the real-time price in 5 min intervals. It also generated an integrated version of real-time price which is the average of 12 set prices in each hour. In this paper, the integrated electricity price is the target variable to predict. in this market. The data set consists of hourly integrated electricity price in real-time of NYC from 01-03-2021 to 28-02-2024, including 36 months. The last twelve months are considered for the test set. Table 2 summarizes the characteristics of this data.

Table 2: Description of integrated real-time electricity price data used in this paper

Dataset	Training samples						Val.	Testing samples					
NYC integrate real-time	1 March 2021 to 28 Feb 2023							1 March 2023 to 28 Feb 2024					
	Max	Min	Std	Skew	Kurt	N	%	Max	Min	Std	Skew	Kurt	N
	3121	-8.1	72.2	16.5	537.3	17520	20	1195	-2.9	32.9	13.6	336.7	8784

[5] www.nyiso.com

2.2 Baseline Models

2.2.1 Statistical model

One of the widely used statistical model for time series forecasting is ARIMA where the three parameters of the model are: the autoregressive parameters $(\phi_t, \dots, \phi_p)$, the moving average parameters $(\theta_t, \dots, \theta_q)$, and $\nabla^d X_t$ which is the lag operator. It is denoted by ARIMA(p ,d, q) where p and q are autoregressive and moving average components respectively and d denotes the number of differencing happens at lag-one (Weron, 2014). The formula can be written as:

$$\phi(B)\nabla^d X_t = \theta(B)\varepsilon_t \tag{1}$$

Where ε_t is the error term at time t, $\phi(B)$ and $\theta(B)$ represent polynomials in the backshift or lag operator.

2.2.2 Machine Learning models

Lasso linear regression is one of the first learning algorithm that revolutionized regression based models (Tibshirani, 1996). The difference between Lasso and regular regression is that it minimizes the residual sum of squares subject to a constraint that makes the sum of the absolute value of the coefficients to be less than a threshold. This regularization approach prevent overfitting through penalizing coefficients.

$$\min \sum_{i=1}^{N} \left(y_i - \sum_j \beta_j \, x_{ij} \right)^2 + \lambda \sum_j |\beta_j| \tag{2}$$

Where N is the total number of observations, X is the predictor, β is the coefficient to estimate, and $\boldsymbol{\lambda}$ is a regularization parameter that controls the strength of the regularization term.

XGB stands for eXtreme Gradient boosting library utilizing regularization on gradient boosting framework (Chen & He, 2015). Gradient boosting builds ensemble model where each weak/base learner, typically a simple decision tree, is trained sequentially to minimize the gradient of a predefined loss function (Friedman, 2001). XGB algorithm can be formulated as the following optimization problem aimed at minimizing a loss function with regularization.

$$\min \sum_{i=1}^{N} \mathcal{L}\left(y_i - \sum_{k=1}^{K} f_k(x_i) \right) + \lambda \sum_{j=1}^{J} \Omega(f_j) \tag{3}$$

Where *N* is the total number of observations, $\mathcal{L}$ is the loss function that measures the discrepancy between the ground truth y_i and the predicted output $\sum_{k=1}^{K} f_k(x_i)$ the i-th observation, f is individual tree (or base learner) in the ensemble, *K* is the total number of trees, *J* is the total number of leaf node, $\Omega(f_j)$ is a regularization term that penalizes the complexity of each individual tree, and λ is a regularization parameter.

Random Forest (RF) is also a tree-based ensemble model widely used for prediction. Each tree is trained independently using a random subset of the training data and a random subset of the input features, overcoming the overfitting issue of decision trees. The generalization error converges as the number of trees increases (Breiman, 2001). The final prediction is obtained by averaging the predictions of all individual trees as the following formula indicates.

$$\hat{y}_i = \frac{1}{K} \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

Where $\hat{y}_i$ represents the predicted output for the i-th observation, f is individual tree, K is the total number of trees, $f_k(x_i)$ is the prediction of the n-th tree for the input features x_i.

Unlike traditional regression, instead of minimizing the error between the predicted and actual values using a chosen loss function, Support vector approach (SVR) aims to minimize the generalization error. It is based on the concept of support vectors (Cortes & Vapnik, 1995), aiming to find a hyperplane in the feature space that best fits the training data while maximizing the margin and minimizing errors. In SVR, the goal is to find the optimal hyperplane that separates the data into two classes: those within a certain margin (epsilon) from the hyperplane, called support vectors, and those outside the margin (Drucker et al., 1996). This optimization problem is formulated to minimize the empirical risk, penalized by a regularization term, subject to the constraints defined by the margin and epsilon-insensitive loss (Smola & Schölkopf, 2004). Mathematically, the prediction $\hat{y}(x)$ in SVR using kernel function can be represented as:

$$\hat{y}(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\, K(x_i, x) + b \tag{5}$$

Where x is the new input data point for which the prediction is made, x_i are the support vectors from the training data, $K(x_i, x)$ is the kernel function, which effectively computes the dot product of the input vectors and new data points in a transformed feature space. α_i and α_i^* are the Lagrange multipliers, which are non-zero only for support vectors and indicate the importance of the corresponding data points, and b is the bias term.

2.2.3 Deep Learning models

Long Short-Term Memory (LSTM) is a form of recurrent neural network (RNN) developed to address vanishing gradient problem in back-propagation process of classic RNNs by introducing specialized memory cells and gating mechanisms that allow the network to selectively remember or forget information over time (Hochreiter & Schmidhuber, 1997).

LSTM networks contain memory cells, which are composed of three main components: the input gate, the forget gate, and the output gate. The input gate controls what piece of new input information to keep in the current state of the memory cell, the forget gate decides which information to abandon from the cell's previous state, and the output gate determines the information to be output from the current state (Gers et al., 2000). Mathematically, the forward direction can be represented as follows.

$$\begin{gathered} f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ g_t = \tanh\left(W_g x_t + U_g h_{t-1} + b_g\right) \\ c_t = f_t \odot c_{t-1} + i_t \odot g_t \\ h_t = o_t \odot \tanh(c_t) \end{gathered} \tag{6}$$

Where at time t, x_t is the input vector, h_t is the hidden state vector, c_t is the cell state vector, $f_t, i_t, o_t,$ and g_t are the forget gate, input gate, output gate, and candidate memory cell activation vectors, respectively. σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication, b_* are bias vectors, W_* and $U_*$are weight matrices for the input and hidden state connections, respectively. The above formulas mechanics can be seen in the figure 1(a).

Gated Recurrent Unit (GRU) is another form of RNN that simplifies the architecture of LSTM by merging the forget and input gates into a single update gate, resulting in fewer parameters and faster training times (Cho, van Merrienboer, Gulcehre, et al., 2014). The mathematical structure is shown the following formula:

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ g_t &= \tanh\left(W_g x_t + U_g (r_t \odot h_{t-1}) + b_g\right) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot g_t \end{aligned} \tag{7}$$

Where at time t, z_t is the update gate, r_t is the reset gate, h_t is the hidden state vector, and g_t is the candidate memory cell activation vectors, respectively. σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication, b_* are bias vectors, W_* and $U_*$are weight matrices for the input and hidden state connections, respectively. GRU architecture is represented in the figure 1(b).

Bidirectional Long Short-Term Memory (Bi-LSTM) is made up of two LSTMs one that processes input in a forward direction and the other in reverse. This allows the Bi-LSTM to monitor information flow from both the previous and following timesteps, which helps it capture long-term dependencies (Schuster & Paliwal, 1997). In other words, in LSTM information is only being carried through a forward layer structure. However, as shown in figure 1, Bi-LSTM adds a backward layer to this structure enabling it to incorporate information both ways.

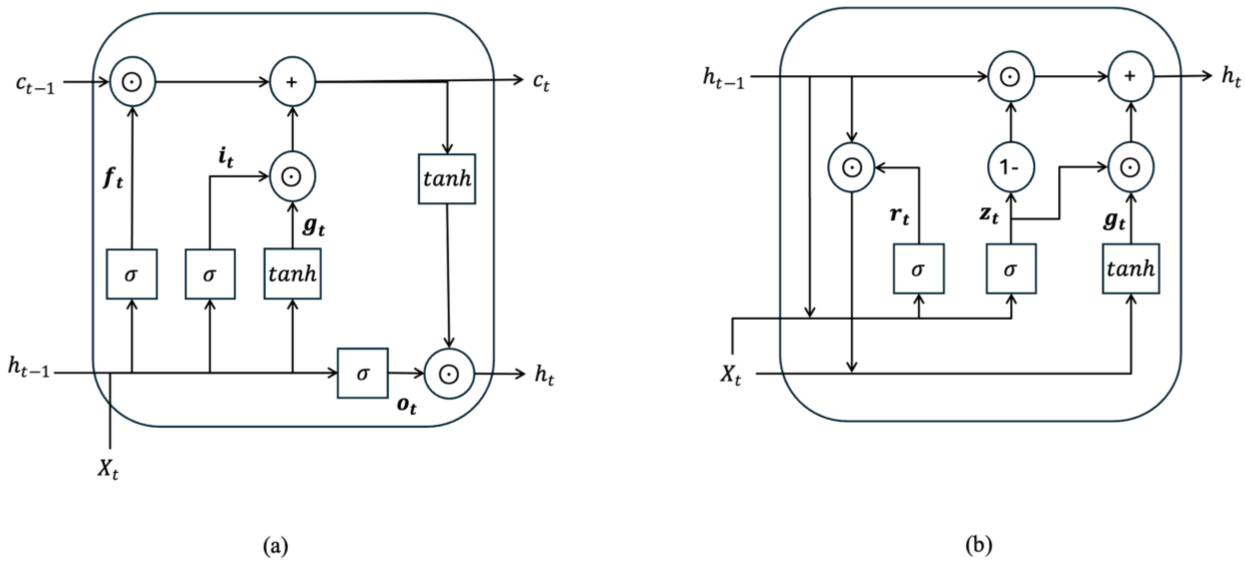

Figure 2: (a) LSTM cell, (b) GRU cell (Olah, 2015)

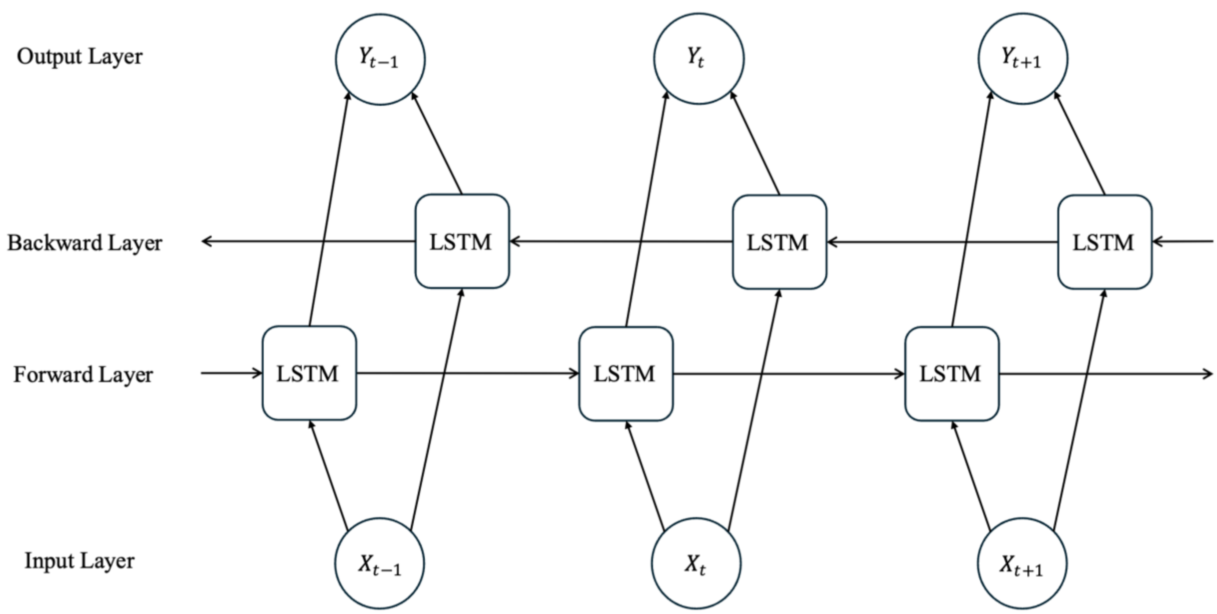

Figure 3: Bi-LSTM architecture (Olah, 2015)

Another class of deep neural networks is Convolutional Neural Networks (CNNs) that automatically learns features and the hierarchical representation of data via filters (or kernel) optimization and addresses vanishing/exploding gradients problems through regularization (Lecun et al., 1998).

A typical CNN architecture consists of an input layer, convolutional layer, pooling layer, and a fully connected layer. The input layer receives the input data with defined shape (length and number of features). Convolutional layer extracts temporal features from the input data by applying convolution operations using learnable filters. Each filter slides (convolves) across the input sequence, computing dot products between the filter weights and the input patch it overlaps (Goodfellow et al., 2016). Mathematical formulation is as follow:

$$(f * x)_t = \sum_{m} x_{t+m} \cdot f_m \tag{8}$$

Where x is the input data, f is the filter, and m represents the position within the filter ranging over the size of the filter, and thus f_m is the filter weights at position m learned during training.

Pooling layer slides a window over the input feature map and takes the maximum value within each window. Through this operation, it reduces the temporal dimensions of the feature maps, leading to a reduction in the number of parameters and computation in the network, and helping to make the detected features invariant to small translations. After several layers of convolutions and pooling, the output is flattened and fed into one or more fully connected layers to compute a weighted sum of the inputs and applies a non-linear activation function (Goodfellow et al., 2016).

2.2.4 Hybrid models

Hybrid models in time series forecasting integrate two or more distinct modeling approaches to leverage their individual strengths and mitigate their weaknesses, aiming for improved prediction accuracy. Initially, statistical methods, such as ARIMA, were combined with ML techniques like SVR leveraging linear capability of statistical models and non-linear capacity of ML architecture (Zhang, 2003). In recent years, however, the focus has shifted towards combining deep learning models to further improve forecasting accuracy. For instance, hybrid models now often integrate CNN with RNN, capitalizing on the CNN's strength in feature extraction and the RNN's ability to capture temporal dependencies (Ehsani et al., 2024; Ghimire et al., 2024). In an experiment within the next section, this paper studies three CNN-RNN hybridizations and compares their forecasting performance with the proposed model.

2.3 Proposed Model

This paper proposes a novel hybrid model combining Bi-LSTM and GRU networks for the task of electricity price forecasting. This hybrid architecture leverages the strengths of both Bi-LSTM and GRU, making it particularly effective for capturing complex temporal dependencies and nonlinear patterns in time series data (Q. Li et al., 2023; X. Li et al., 2022; Michael et al., 2024), which are inherent in price related data (Althelaya et al., 2018; Karim et al., 2022).

The model begins with a BiLSTM layer. BiLSTM networks are well-known for their capability to learn long-term dependencies in sequence data due to their ability to retain information over long periods. The bidirectional nature of the BiLSTM allows the model to learn in backward and forward fashion, providing a more comprehensive understanding of the temporal dynamics (Graves et al., 2013; Schuster & Paliwal, 1997). This is crucial for identifying hourly electricity price patterns. A dropout layer is added after the Bi-LSTM layer to prevent overfitting. Dropout is a regularization technique that randomly drops units (along with their connections) during training, which helps in making the model more robust (Srivastava et al., 2014). Following the dropout, a GRU is employed. It adds another layer of sequential analysis to capture significant temporal dependencies yet in a more efficient way due to fewer parameters compared to LSTMs (Cho et al., 2014). Another dropout layer is included after the GRU layer to further ensure robustness against overfitting. Finally, a dense (fully connected) layer with a single neuron is used to produce the output, which in this case is the forecasted electricity price.

The implementation of the Bi_LSTM-GRU is carried out using the Keras library. The dataset used for training and evaluation is normalized to ensure that all features contribute equally to the training process. The sequences were generated using a sequence length of 24, which corresponds to 24 hours, aligning with the granularity of hourly electricity price data. Figure 3 shows the architecture of the proposed hybrid DL model. It can be generalized to other DL models through modifying the modeling block.

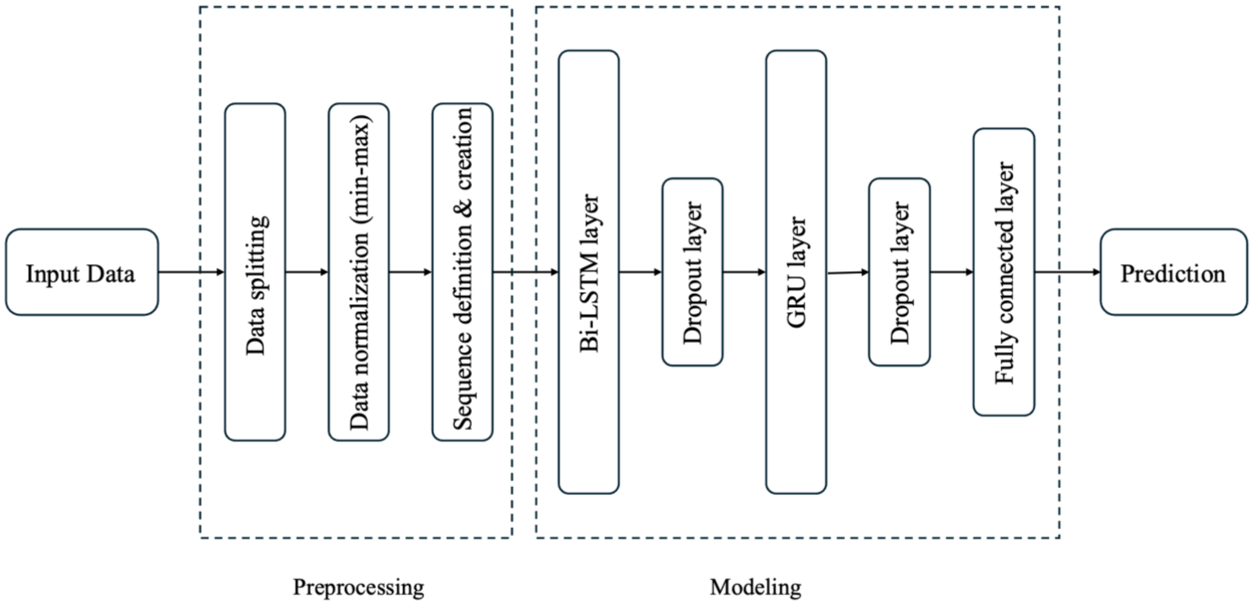

Figure 4: Architecture of proposed Bi-LSTM-GRU hybrid model

2.3.1 Rationales for using Bi-LSTM-GRU for electricity price forecasting

Hybridizing Bi-LSTM with GRU for electricity price forecasting offers several unique advantages that may not be as effectively achieved by other models or different combinations of neural network architectures. Three key merits of such hybrid model are:

I. **Efficient Deep Learning**: Capturing extensive contextual dependencies typically needs a deep neural network, which can significantly increase computational resources and training time. The proposed model strategically balances in-depth learning with computational efficiency. Complementing Bi-LSTM with a simpler, yet deep GRU model effectively enhances accuracy by leveraging a deeper network architecture while mitigating the computational demands associated with increased depth.

II. **Superior Error Correction**: In the hybrid Bi-LSTM and GRU model, error signals are back-propagated first through the GRU for initial feature filtering adjustments, then through the Bi-LSTM for comprehensive refinement using its bidirectional structure. This sequential and layered approach to error correction significantly enhances the model's precision in tuning predictions, making it exceptionally effective in complex and noisy environments such as electricity price markets. This hybrid method ensures robust convergence on accurate predictions despite the complex temporal dynamics involved.

III. **Enhanced Robustness to Sequence Length Variability**: In the electricity market, factors like weather can have unpredictable effects on demand and supply, often deviating from typical daily patterns. To effectively accommodate these variations and maintain forecasting accuracy, this hybrid model first leverages the Bi-LSTM's gate mechanisms to dynamically adjust its sensitivity to former and later data points. This allows the model to effectively respond to changes in sequence length triggered by such conditions. Subsequently, the GRU layer focuses on refining the most crucial features extracted by the Bi-LSTM, focusing on elements that are predictive of subsequent price points. This ensures robust prediction, even as data characteristics fluctuate significantly.

2.4 Data decomposition techniques

Decomposition techniques can be employed in data preprocessing stage of time series forecasting, as they decompose the original, non-stationary data with high fluctuation into more stable subseries, facilitating better analysis and prediction. In recent years, these techniques have been widely applied to energy forecasting with favorable outcomes (Chang et al., 2019; Qiao & Yang, 2020). Typically, decomposition-based hybrid models follow a consistent framework: the original time series is segmented into various subseries; each subseries is then forecasted separately; finally, an aggregate prediction is formulated by combining the results from all subseries (Liu & Chen, 2019). The Fourier transform (FT) is a fundamental techniques in signal analysis, widely used in energy forecasting to address non-stationary time-series (González-Romera et al., 2008; Yu et al., 2018). FT converts a time series from the time domain to the frequency domain, allowing for the identification of individual frequency components. This provides a comprehensive view of the global frequency content of the signal (Musbah & El-Hawary, 2019). The discrete Fourier transform (DFT) is as follow:

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-2\pi i.\frac{k.n}{N}} \quad \text{for} \quad k = 0, 1, 2, \ldots, N-1 \tag{10}$$

Where X_k represents the *k-th* frequency component of the original sequence, x_n is the original time-domain signal at the *n-th* time point, N is the total number of points in the sequence, $e^{-2\pi i.\frac{k.n}{N}}$ is the complex exponential function that encodes the frequency information, and i is the imaginary unit.

Wavelet transform (WT) is another frequency-based decomposition method that enhances the capabilities of FT by enabling time-frequency localization and multiresolution analysis (Yao et al., 2000). When combined with neural network models, WT has significantly improved prediction accuracy in various forecasting tasks, including generation forecasting (Liu et al., 2013), load forecasting (Rana & Koprinska, 2016), and price forecasting models (Chang et al., 2019; Qiao & Yang, 2020). WT excels in analyzing signals with transient features and localized variations by facilitating a multi-resolution analysis that captures both time and frequency dimensions. It decomposes the original time series using low-frequency and high-frequency filters. The low-frequency filter generates an approximate series, while the high-frequency filter produces detailed series. Further decompositions of the low-frequency series at each level result in a final output consisting of one low-frequency approximate subseries and multiple high-frequency detailed subseries. Both the approximation and detail coefficients are used for training and testing (Qian et al., 2019). Below, the mathematical description of Wavelet Transform (WT) for discrete series is provided:

$$x(t) = \sum_{k} a_k . \phi_k(t) + \sum_{j=1}^{J} \sum_{k} d_{j,k} . \psi_{j,k}(t) \tag{9}$$

Where $\phi_k(t)$ is the scaling function, which captures the low-frequency (approximation) components of the signal, a_k are the approximation coefficients, $\psi_{j,k}(t)$ is the wavelet function, which captures the high-frequency (detail) components at different scales, $d_{j,k}$ are the detail coefficients at scale j and position k.

This paper investigates the impact of frequency-based decomposition methods during the preprocessing stage, particularly in the context of electricity price forecasting, where periodic patterns, cycles, and seasonality are often present. The Fast Fourier Transform (FFT), an optimized algorithm to compute the DFT, is employed for its efficiency in transforming data. Similarly, the Discrete Wavelet Transform (DWT) is selected from the WT family for its effectiveness. The decomposition level parameter in DWT, which controls the number of iterative decompositions into lower resolution components, is set to 5 to balance capturing fine details and computational complexity. The 'db1' wavelet is chosen for its simplicity and efficiency in capturing abrupt changes in signal levels (Daubechies, 1992).

As shown in Figure 4, decomposition techniques are typically applied after data normalization. This step ensures numerical stability and consistent interpretation, which is especially important when dealing with data that exhibits large variations in magnitude (Chang et al., 2019). The decomposition function outputs a list of coefficients/components, which are then used to create sequences that are fed into the prediction model.

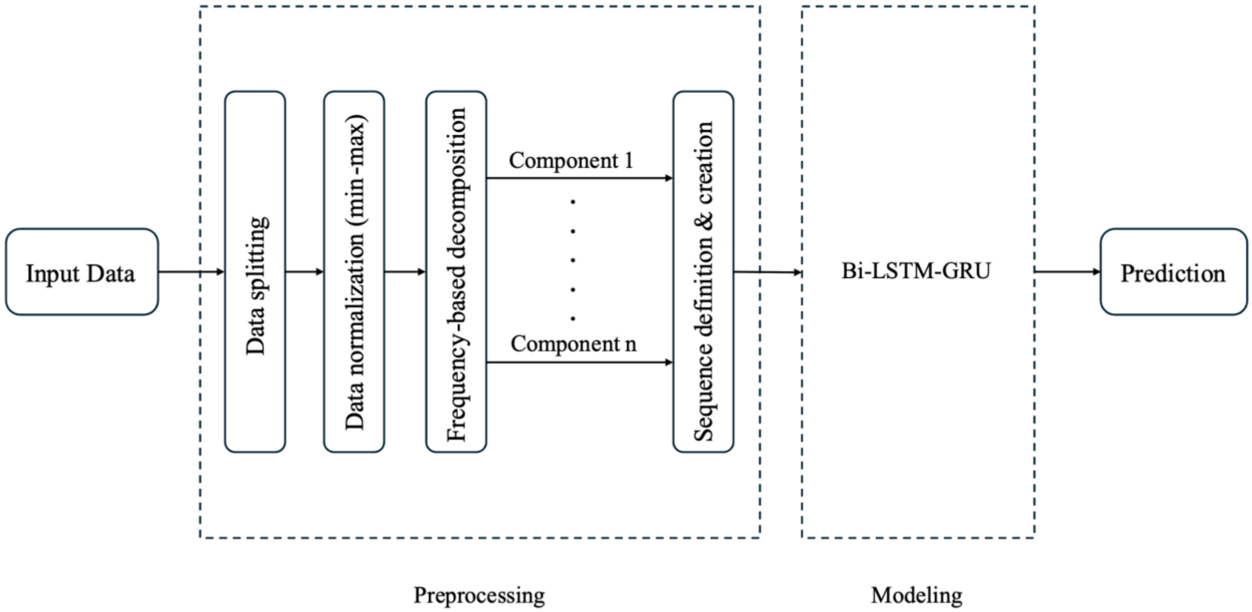

Figure 5: Data decomposition stage in preprocessing

2.5 Evaluation criteria

Systematic comparison of the prediction performance of different models in the domain of energy price forecasting can be conducted utilizing various evaluation metrics (Lu et al., 2021). Two widely used metrics are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE measures the average magnitude of errors between predicted and actual values, providing a straightforward interpretation of the forecast accuracy. It is calculated using the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{10}$$

where $\boldsymbol{y_i}$ represents the actual values, $\boldsymbol{\hat{y}_i}$ represents the predicted values, and n is the number of observations. MAE is robust to outliers, as it does not square the errors, making it particularly useful when large deviations are not as critical (Willmott & Matsuura, 2005).

In contrast, RMSE gives more weight to larger errors due to the squaring of the error terms, making it sensitive to outliers. RMSE is particularly useful when large errors are highly undesirable and need to be penalized more severely. This metric provides a measure of the standard deviation of the residuals (prediction errors), giving insight into the model's overall error distribution (Chai & Draxler, 2014). Following formula shows its calculation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

Mean absolute percentage error (MAPE) is another widely used metric in energy market, especially for demand or load forecasting. However, it is not applicable for EPF due to an undefined formula in case of having a zero price (Ehsani et al., 2024).

3 Empirical Results

The models described in the previous section of this paper are trained and tested on this data separately. Table 3 indicates the value of hyperparameters for all models as well as the explanatory variables used. The hyperparameter space is defined based on relevant models from the literature, and grid search is employed to optimize hyperparameter selection in both statistical and machine learning models. The hyperparameter space is defined based on relevant models from the literature, and grid search is employed to optimize hyperparameter selection in both statistical and machine learning models. According to the literature, hyperparameter tuning has a relatively minor effect on the performance compared to the architecture of deep learning models. Therefore, given the high computational demands of deep learning, this study prioritizes comparing different architectures while keeping the same hyperparameters consistent across all models.

Table 3: Summary of the models hyperparameters and variables

Model	Hyperparameters	Variables
ARIMA	(5, 1, 1)	
Lasso	alpha = 7	Temporal dummies, Lag24
XGB	colsample_bytree: 0.840, gamma: 0.354, learning_rate: 0.016, max_depth: 4, min_child_weight: 8.219, n_estimators: 513, reg_alpha: 0.212, reg_lambda: 0.181, subsample: 0.673	Temporal dummies, Lag24
RF	Bootstrap: True, max_depth: 6, max_features: 'auto', min_samples_leaf: 8, min_samples_split: 8, n_estimators: 221	Temporal dummies, Lag24
SVR	C=100, epsilon=0.1, kernel='rbf'	Temporal dummies, Lag24
LSTM	LR = 0.001, units = 50, Dense layer = 1, $optimizer = Adam$	None
Bi-LSTM	LR = 0.001, units = 50, Dense layer = 1, $optimizer = Adam$	None
GRU	LR = 0.001, units = 50, Dense layer = 1, $optimizer = Adam$	None
CNN	LR = 0.001, filters = 64, , $kernel\_size$ = 3, $optimizer = Adam$	None
Bi-LSTM-GRU	LR = 0.001, units = 50, Dense layer = 2, Dropout=0.2, $optimizer$ = N$Adam$	None

The forecasting performance of baseline models being measured in terms of MAE and RMSE metrics and being compared against the proposed model performance are reported in Table 4. The proposed model shows the lowest error among others.

Table 4: Results of forecasting models on Integrated Realtime Market electricity price in NYC

Model	Statistical	Machine Learning (ML)				Deep Learning (DL)					
Metric	ARIMA	LR (Lasso)	XGB	RF	SVR	CNN	LSTM	Bi-LSTM	GRU	GRU-Bi_LSTM	Bi_LSTM-GRU (Proposed)
RMSE	38.745	45.318	48.838	48.955	39.279	25.238	24.175	24.773	24.143	24.781	23.604
MAE	27.876	35.653	35.989	36.058	26.535	8.125	7.708	6.809	6.848	7.329	6.602

Figures 5 depicts the graphs of actual price and the predicted price by the proposed model, over the period of the test set from March 2023 to March 2024 in NYC. The graphs of the forecast result of the ML and DL baseline models versus true values are represented in figures 6 to 14.

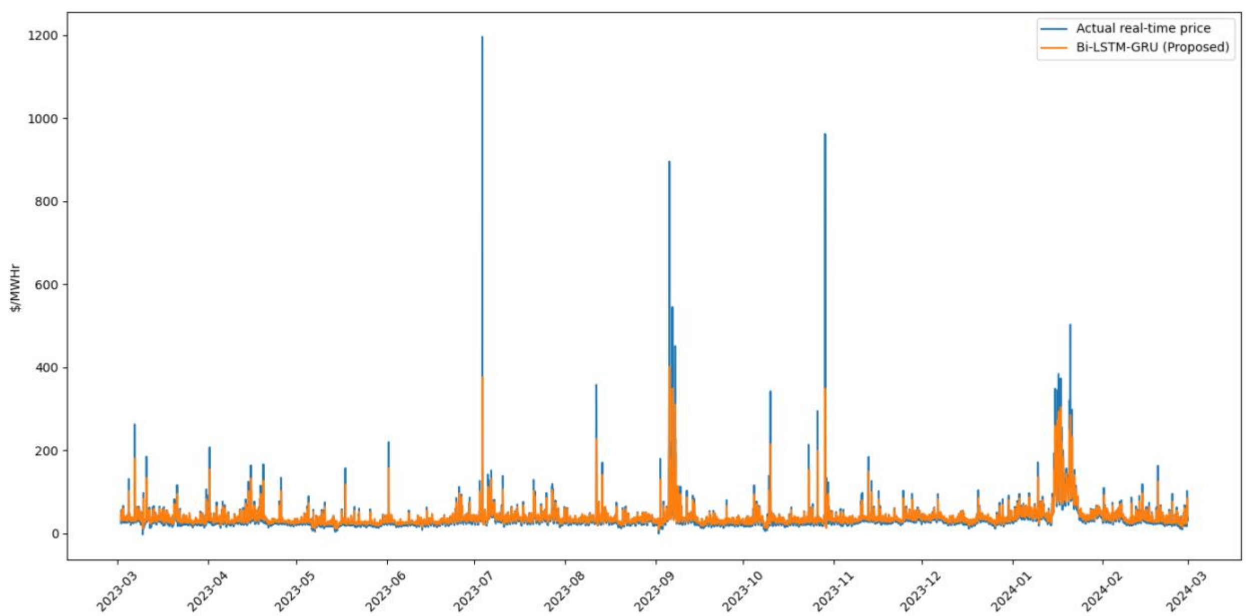

Figure 6: Hourly electricity price for NYC from March 2023 to March 2024 and the proposed Bi-LSTM-GRU model prediction

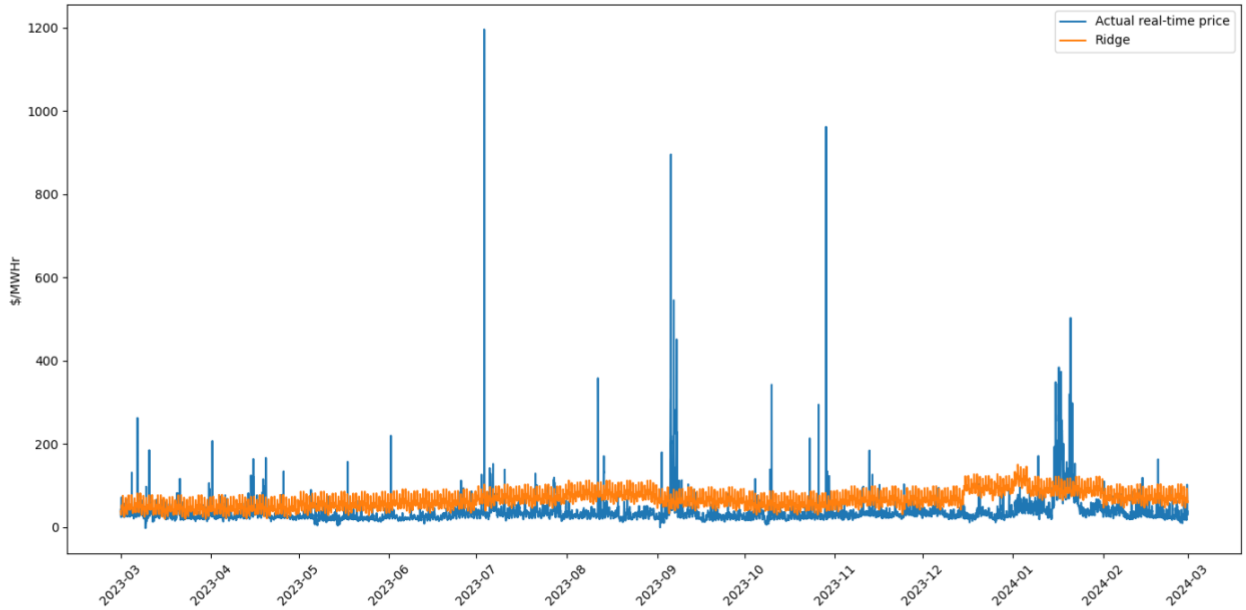

Figure 7: Hourly electricity price for NYC from March 2023 to March 2024 and Ridge model prediction

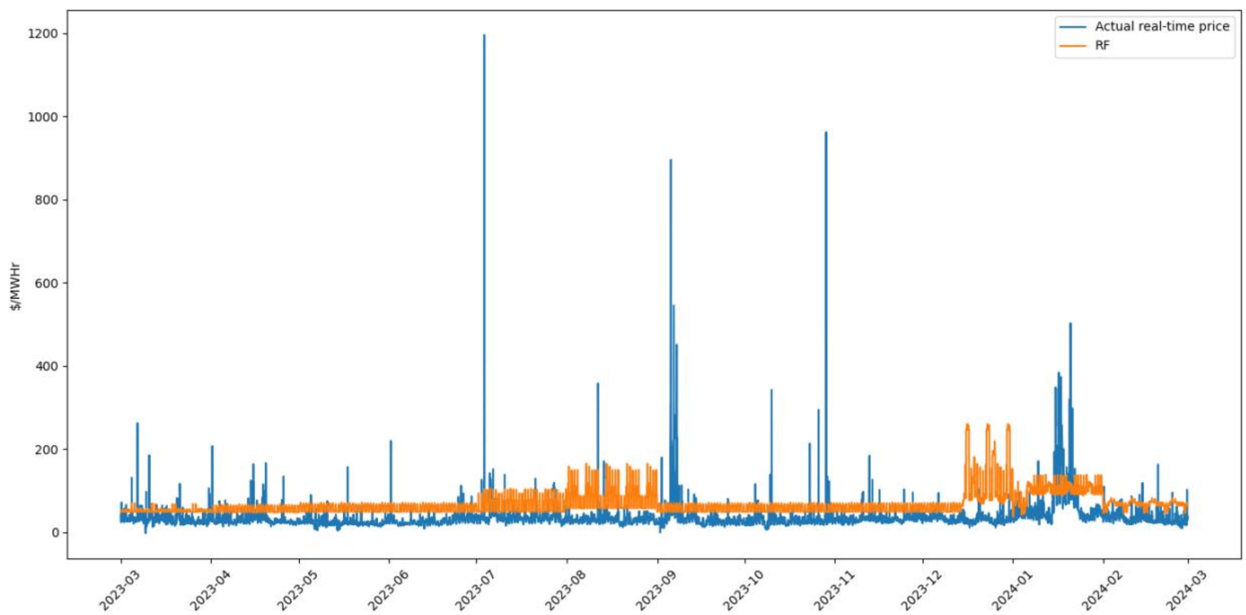

Figure 8: Hourly electricity price for NYC from March 2023 to March 2024 and RF model prediction

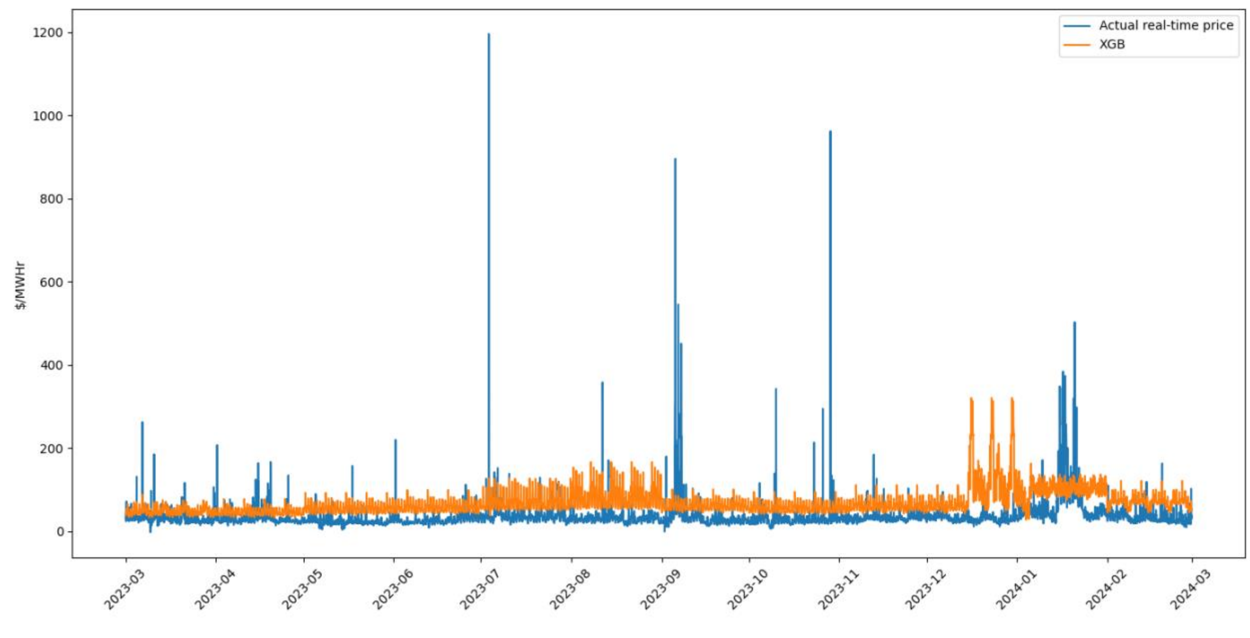

Figure 9: Hourly electricity price for NYC from March 2023 to March 2024 and XGB model prediction

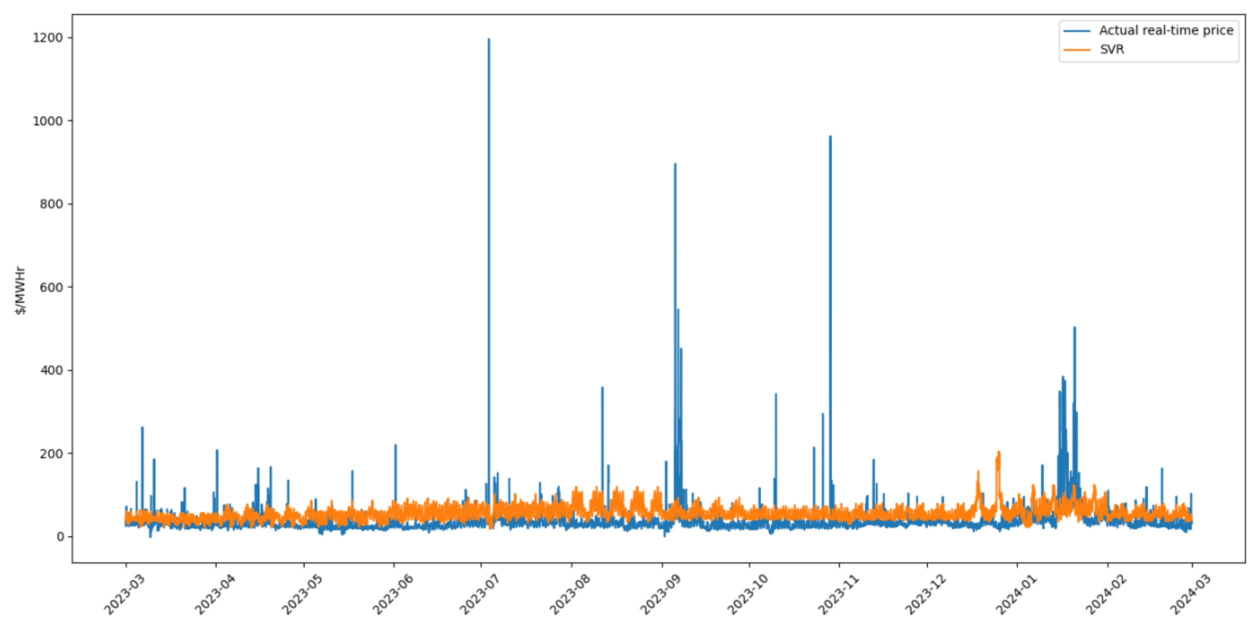

Figure 10: Hourly electricity price for NYC from March 2023 to March 2024 and SVR model prediction

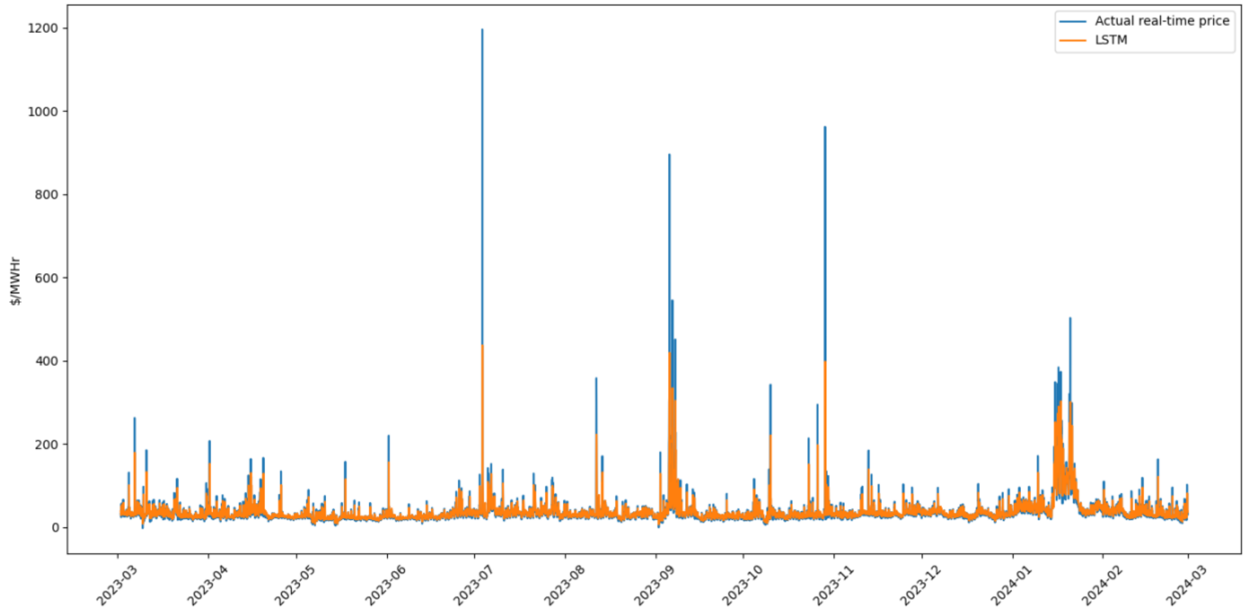

Figure 11: Hourly electricity price for NYC from March 2023 to March 2024 and LSTM model prediction

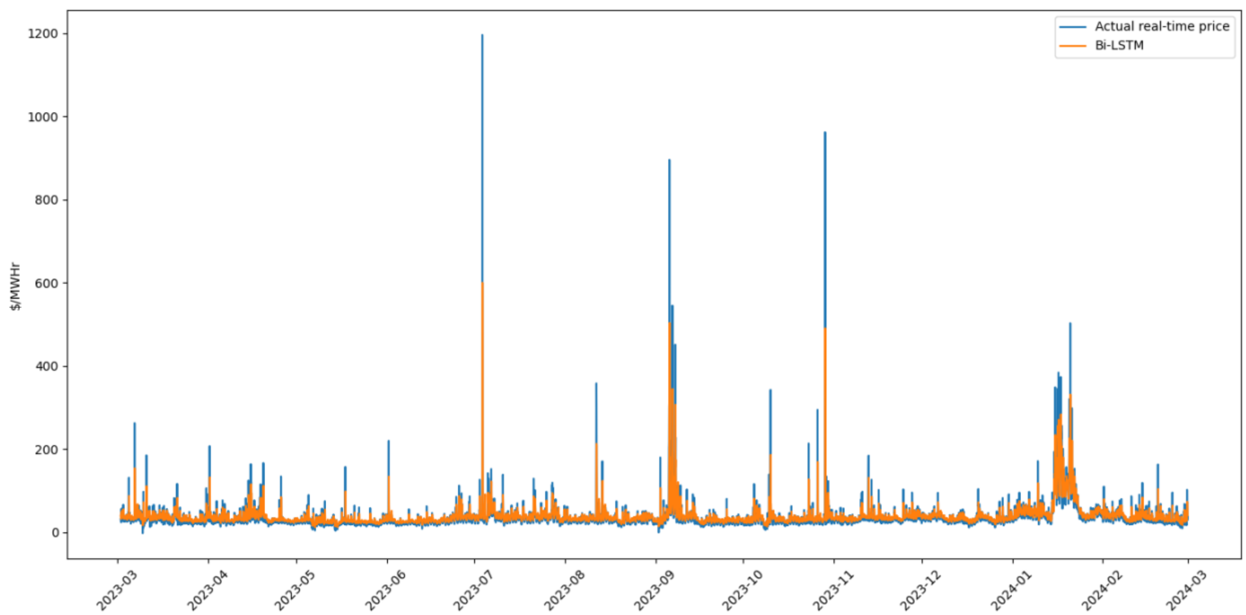

Figure 12: Hourly electricity price for NYC from March 2023 to March 2024 and Bi-LSTM model prediction

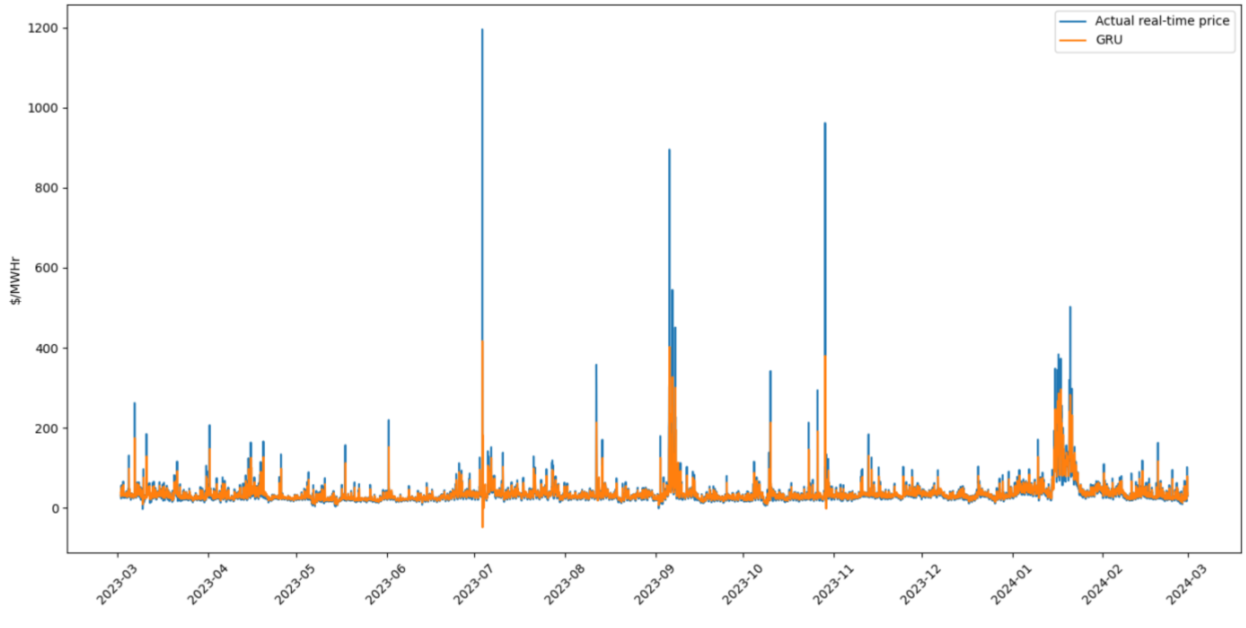

Figure 13: Hourly electricity price for NYC from March 2023 to March 2024 and GRU model prediction

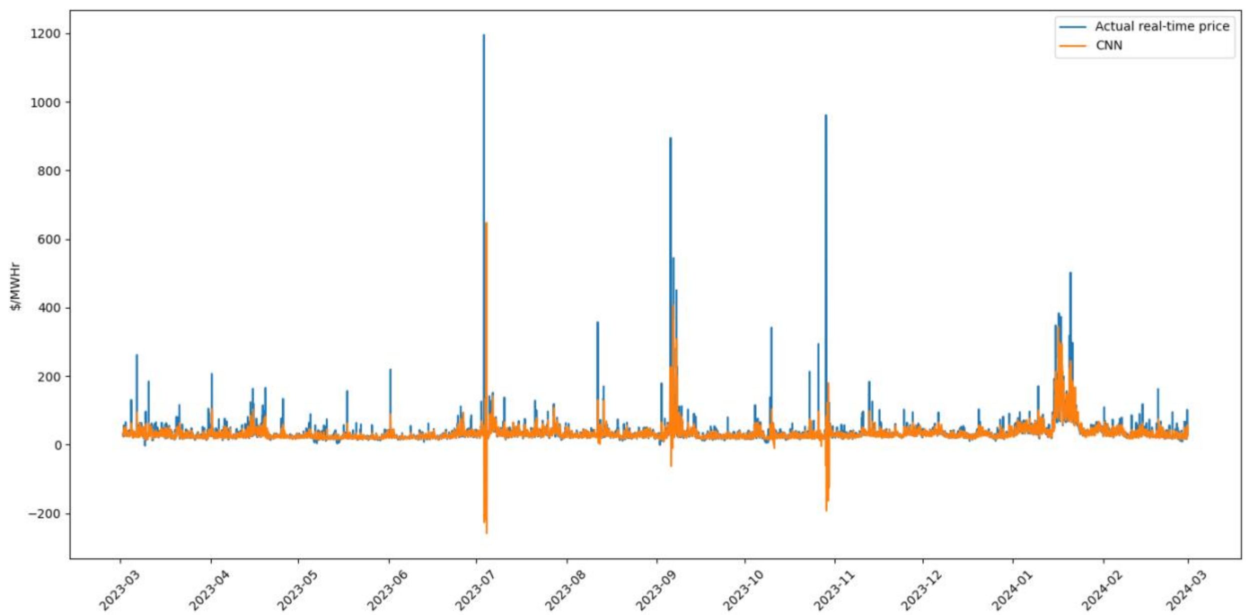

Figure 14: Hourly electricity price for NYC from March 2023 to March 2024 and CNN model prediction

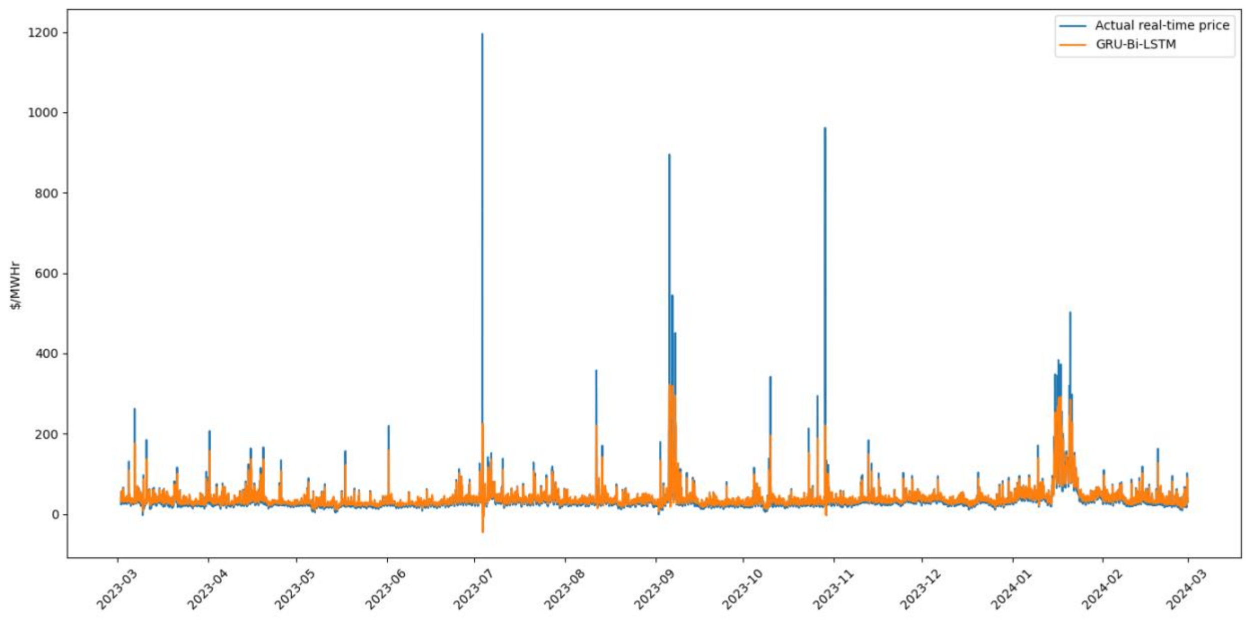

Figure 15: Hourly electricity price for NYC from March 2023 to March 2024 and GRU-Bi-LSTM model prediction

3.1 CNN-RNN Experiment

In this experiment, the possible contribution of combining CNN architecture with that of RNN is explored. Three CNN-RNN models are trained and tested on the data and the performance then compared with the proposed model. As shown in Table 5, adding CNN is of limited contribution and none of the new hybrid models can beat the proposed model performance.

Table 5: Comparison of CNN-RNN hybridization with the proposed model

Metric	CNN-LSTM	CNN-Bi_LSTM	CNN-GRU	Bi-LSTM-GRU (Proposed)
RMSE	24.019	24.367	24.402	23.604
MAE	7.068	7.368	7.551	6.602

Figures 15 to 18 represent the graphs of the forecast result of the models in this experiment.

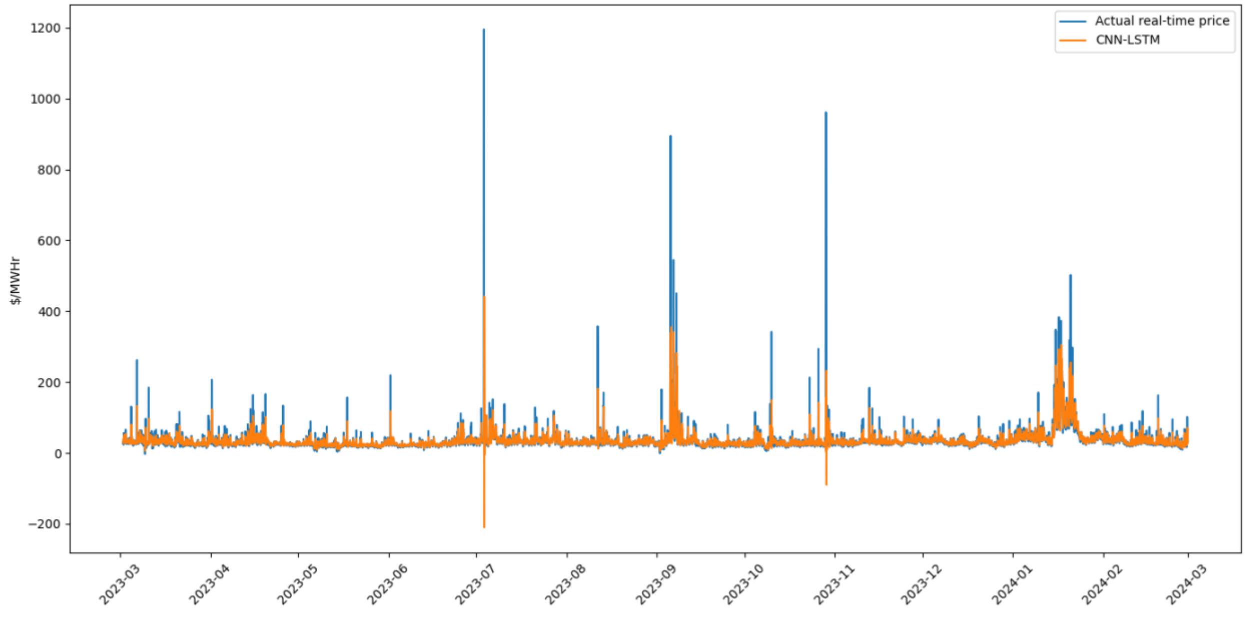

Figure 16: Hourly electricity price for NYC from March 2023 to March 2024 and CNN-LSTM model prediction

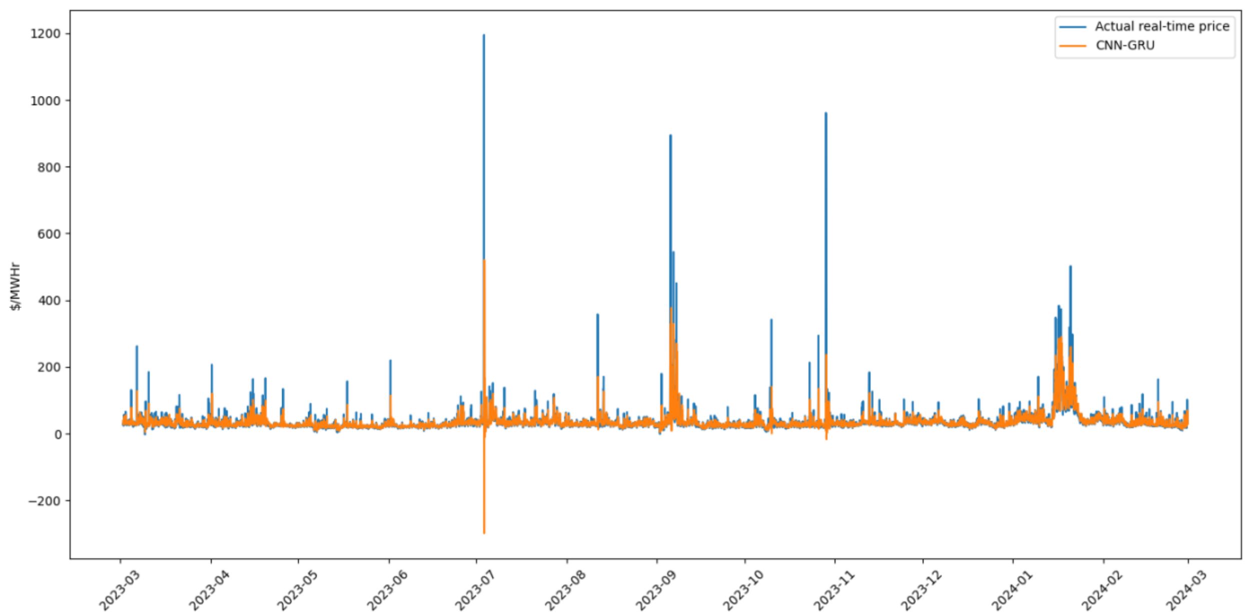

Figure 17: Hourly electricity price for NYC from March 2023 to March 2024 and CNN-GRU model prediction

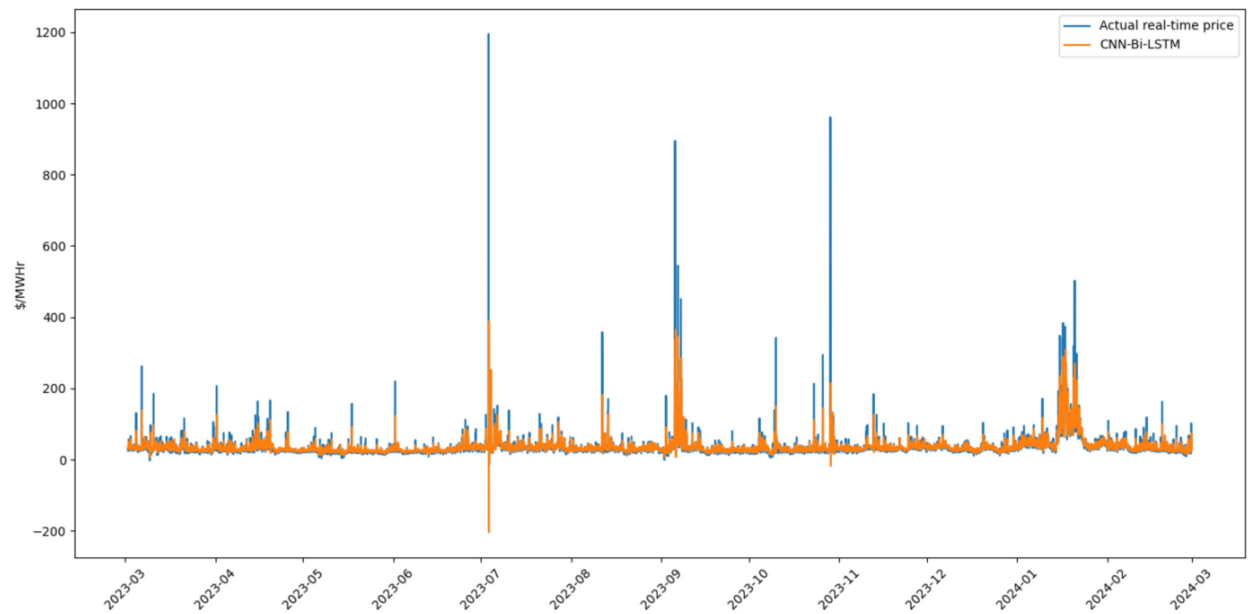

Figure 18: Hourly electricity price for NYC from March 2023 to March 2024 and CNN-Bi-LSTM model prediction

3.2 Decomposition Experiment

This experiment is designed to answer the question of whether data decomposition can further enhance the proposed model. WT and FT techniques are employed in data preprocessing stage before applying the proposed model for prediction. The results in table 6 indicate that the proposed model will be worst off after employing these two data decomposition techniques.

Table 6: The effect of adding a data decomposition method to the proposed model

Metric	WT-Bi-LSTM-GRU	FT-Bi-LSTM-GRU	Bi-LSTM-GRU (Proposed)
RMSE	30.734	23.680	23.604
MAE	10.261	6.702	6.602

Electricity prices are influenced by a variety of factors, including demand and supply dynamics, weather conditions, market regulations, and random shocks. While some of these factors exhibit periodic or cyclic patterns, which frequency-based decomposition might capture, the data can also contain a significant amount of noise, abrupt changes, or irregular trends that are not well-suited to this type of decomposition, leading to inferior performance. FT may not significantly alter the model's performance because it is capturing the dominant frequency components that are already well-represented in the raw data. This suggests that the model is already capturing these components effectively without the need for additional transformation. WT decomposes the signal into approximation and detail coefficients. If the detail coefficients do not add meaningful information or if they introduce noise, the model's performance can degrade. Specifically, if the high-frequency components captured by WT do not correspond to relevant patterns in the electricity price data but instead capture noise, this can negatively affect the model.

3.3 Discussion

This study compared the performance of various predictive models against the proposed Bi_LSTM-GRU hybrid model for electricity price forecasting. The experimental results clearly indicate that the Bi_LSTM-GRU model outperforms all the other models in terms of widely used evaluation metrics in the domain. In addition to these models, two experiments were conducted: first, a CNN-RNN hybrid model including CNN-LSTM, CNN-GRU, and CNN-Bi-LSTM was built and evaluated, and second, data decomposition techniques such as Wavelet Transform (WT) and Fast Fourier Transform (FFT) were employed alongside the proposed model. Both approaches failed to surpass the performance of the Bi_LSTM-GRU model. Below, we discuss the reasons behind the inferior performance of each alternative model and why our proposed model excels.

ARIMA model performed the worst among all the models evaluated. This is largely due to its linear nature and assumption of stationarity. Since electricity price data often exhibit non-linear patterns and abrupt changes, linear models perform poorly on this task. LASSO also suffers from the same issue as it assumes a linear relationship between the input features and the target variable. Moreover, Lasso's regularization can overly penalize important predictors, leading to underfitting and poor generalization performance.

While XGB, RF, and SVR can handle non-linearity, they do not inherently model the sequential nature of time series data. They require manual feature engineering to create lagged variables and other temporal features. Even with these enhancements, they lack the capability to capture complex temporal patterns and long-term dependencies effectively.

CNNs can capture local patterns through convolutional filters but lack the mechanism to model long-term dependencies and sequential patterns effectively. This limitation becomes evident in the context of electricity price forecasting, where understanding long-term trends and temporal dependencies is crucial. While CNNs performed better than ARIMA, XGBoost, RF, and SVR due to their ability to extract complex features from the data, they still fell short of RNN family models which are specifically designed to capture sequential dependencies. LSTM, Bi-LSTM, and GRU are designed to capture long-term dependencies in sequential data. However, when used individually, they are prone to overfitting.

In the first experiment, various CNN-RNN hybrid models (CNN-LSTM, CNN-GRU, and CNN-Bi-LSTM) were evaluated. While these models benefit from CNN's ability to extract local features and RNN's capability to capture temporal dependencies, they still fell short of the Bi_LSTM-GRU model. The primary reason is that the Bi_LSTM-GRU model's architecture, starting with a bidirectional LSTM layer, allows for more comprehensive temporal feature extraction before the GRU layer refines this information, providing better overall performance.

In the second experiment, data decomposition techniques such as WT and FFT were employed alongside the proposed model. These techniques aim to break down the time series data into different frequency components to improve forecasting accuracy. However, they did not enhance the performance of the Bi_LSTM-GRU model. The likely reason is that while decomposition can provide insights into different components of the data, the Bi_LSTM-GRU model already captures complex temporal dependencies effectively, making additional decomposition redundant or even disruptive.

The proposed Bi_LSTM-GRU model significantly outperformed baseline models due to its ability to fully leverage the strengths of both Bi-LSTM and GRU architectures. The Bi-LSTM layer first captures comprehensive temporal dependencies by processing the input sequence in both forward and backward directions. This enriched representation is then refined by the GRU layer, which processes the sequential data efficiently, ensuring that important temporal patterns are retained without excessive computational overhead. Moreover, introducing dropout layers after the Bi-LSTM and GRU layers helps in regularizing the model by randomly setting a fraction of input units to zero at each update during training time. This prevents the model from overfitting to the training data, thereby improving its generalization capability to unseen data.

Additionally, Bi_LSTM-GRU also outperformed GRU-Bi_LSTM architecture with the same dropout layer structure. This can be explained by the nature of the data and task in this research. Electricity prices are influenced not only by past trends but also by future market conditions, regulations, and external factors. Starting with a Bi-LSTM ensures that the model extracts a comprehensive set of features that incorporate both past and future information. This rich representation is then efficiently processed by the GRU, making the model robust and effective in capturing long-term dependencies without excessive computational overhead.

It is important to acknowledge that direct comparisons between the performance of the proposed Bi-LSTM-GRU model and those reported in the literature are challenging due to differences in datasets, such as the time periods, data resolutions, and geographical focuses. However, it is observed in the literature that RNN models, which belong to the deep learning family, tend to perform best in complex time-series forecasting tasks, especially with highly volatile data such as electricity prices. Hybridization of different deep learning models has proven to be more effective, as this approach can mitigate the shortcomings of individual models, providing greater reliability and flexibility. Nevertheless, there is no universally optimal hybridization strategy; in each context, various scenarios must be explored to determine which approach works best.

3.4 Limitations

Despite the promising results of the proposed Bi_LSTM-GRU model for electricity price forecasting, this model, and this study for that matter, have limitations that must be acknowledged to provide a comprehensive understanding of the findings and implications.

On the modeling side, the proposed model has the common limitations of general DL architecture. It is computationally intensive due to the high number of parameters to calculate and extensive hyperparameter to tune. The model's performance is highly dependent on the quality and quantity of input data, making it vulnerable to degradation when faced with sparse, noisy, or small amount of data. Its "black-box" nature poses challenges for interpretability, which can limit its usage in a company where transparency and explainability of the prediction task are required.

The study's limitations include the three-year period of the data used spanning from 2021 to 2024. This period selection was unavoidable due to the impact of COVID-19 on the electricity price before 2021. Data aggregation from 5-minute intervals to hourly averages may result in a loss of granularity and potentially valuable information. The geographic and market specificity of the data limits the generalizability of the findings to other regions or markets. Additionally, the study does not explicitly account for external factors such as regulatory changes and macroeconomic conditions, which can impact the price in the long run.

3.5 Implications for the electricity market

This study has significant implications across various stakeholders in the electricity market. By providing accurate and reliable price forecasts for the upcoming year, this model can influence decision-making processes, operational strategies, and financial planning.

For electricity generators, precise price forecasting is crucial for optimizing the generation schedule and maximizing profitability. With advanced notice of potential price fluctuations, generators can adjust their output to match periods of high prices, thereby increasing revenue. Additionally, accurate forecasting helps generators make informed decisions regarding fuel purchases, maintenance scheduling, and potential investments in new technologies or capacity expansions.

System operators, who are for ensuring the reliable supply of electricity, can enhance grid reliability through better resource allocation, demand response strategies, and contingency planning. Furthermore, accurate price predictions assist in the integration of renewable energy sources, which can be unpredictable due to their dependence on weather conditions. The ability to forecast electricity prices accurately aids in balancing supply and demand dynamically, maintaining grid stability, and minimizing the cost of electricity for all users.

For wholesale customers such as large industries who buy electricity on the wholesale market or through bilateral contracts, the ability to forecast electricity prices accurately is essential for budgeting, securing the best possible rates and negotiating long-term contracts. Wholesale customers can better plan their energy purchases to take advantage of lower prices and avoid periods of high prices, thus reducing their overall energy costs.

4 Conclusion

This research addressed the complex challenge of electricity price forecasting in the deregulated electricity market of New York, characterized by high volatility, non-stationarity, and multi-seasonality. By developing a novel hybrid deep learning model that combines Bi_LSTM with GRU, the study advanced the field of electricity price forecasting. The Bi_LSTM-GRU model not only mitigates the risk of overfitting but also enhances both the robustness and adaptability of the forecasts, proving particularly effective for real-time application in dynamic market environments.

The performance of the Bi_LSTM-GRU model was thoroughly evaluated against a comprehensive array of both traditional and contemporary forecasting models. Demonstrating superior accuracy, the hybrid model outperformed baseline models including ARIMA, Linear Regression, Random Forest, eXtreme Gradient Boosting, Support Vector Regression, as well as advanced deep learning approaches such as LSTM, Bi-LSTM, GRU, and Convolutional Neural Network. This superiority highlights the efficacy of integrating LSTM and GRU architectures to capture the intricate dynamics of electricity prices more effectively than single-mechanism models.

Further explorations within this study assessed the integration of CNN with RNN architectures and the impact of employing data decomposition techniques like Wavelet Transform and Fourier Transform. The findings revealed that while CNN-RNN hybrids did not surpass the Bi_LSTM-GRU model, the application of data decomposition methods resulted in increased error rates, suggesting that for the NYC electricity market, sophisticated deep learning configurations without data decomposition offer the most accurate forecasting tools.

Future research directions could explore integrating additional data types, such as weather-related data and economic indicators, and testing advanced neural network architectures like Transformers. Another potential direction could be examining the effectiveness of wavelet function type and level of wavelet decomposition on performance of deep learning systems. Additionally, alternative decomposition techniques beyond frequency-based methods could be investigated to capture diverse data patterns and enhance forecasting accuracy. Enhancing model interpretability, reducing computational demands, and developing real-time adaptive learning models that dynamically adjust to new data are also promising areas. These efforts could further refine the theoretical and practical aspects of forecasting models, broadening their applicability and effectiveness in the energy sector.

References

Aggarwal, S. K., Saini, L. M., & Kumar, A. (2009). Electricity price forecasting in deregulated markets: A review and evaluation. *International Journal of Electrical Power & Energy Systems*, *31*(1), 13–22. https://doi.org/10.1016/j.ijepes.2008.09.003

AL-Musaylh, M. S., Deo, R. C., Li, Y., & Adamowski, J. F. (2018). Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. *Applied Energy*, *217*, 422–439. https://doi.org/10.1016/j.apenergy.2018.02.140

Althelaya, K. A., El-Alfy, E.-S. M., & Mohammed, S. (2018). Stock Market Forecast Using Multivariate Analysis with Bidirectional and Stacked (LSTM, GRU). *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 1–7. https://doi.org/10.1109/NCG.2018.8593076

Antonini, M., Barlaud, M., Mathieu, P., & Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Transactions on Image Processing*, *1*(2), 205–220. https://doi.org/10.1109/83.136597

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chang, Z., Zhang, Y., & Chen, W. (2019). Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. *Energy*, *187*, 115804. https://doi.org/10.1016/j.energy.2019.07.134

Chen, T., & He, T. (2015). *xgboost: eXtreme Gradient Boosting*.

Cheng, H., Ding, X., Zhou, W., & Ding, R. (2019). A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. *International Journal of Electrical Power & Energy Systems*, *110*, 653–666. https://doi.org/10.1016/j.ijepes.2019.03.056

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches* (arXiv:1409.1259). arXiv. https://doi.org/10.48550/arXiv.1409.1259

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (arXiv:1406.1078). arXiv. https://doi.org/10.48550/arXiv.1406.1078

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics.

Dragomiretskiy, K., & Zosso, D. (2014). Variational Mode Decomposition. *IEEE Transactions on Signal Processing*, *62*(3), 531–544. https://doi.org/10.1109/TSP.2013.2288675

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support Vector Regression Machines. *Advances in Neural Information Processing Systems*, *9*. https://proceedings.neurips.cc/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html

Ehsani, B., Pineau, P.-O., & Charlin, L. (2024). Price forecasting in the Ontario electricity market via TriConvGRU hybrid model: Univariate vs. multivariate frameworks. *Applied Energy*, *359*, 122649. https://doi.org/10.1016/j.apenergy.2024.122649

FERC. (2024). *Energy Primer: A Handbook of Energy Market Basics | Federal Energy Regulatory Commission*. https://www.ferc.gov/news-events/news/ferc-staff-issues-2024-energy-primer-handbook-energy-market-basics

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. Neural Computation. https://doi.org/10.1162/089976600300015015

Ghimire, S., Deo, R. C., Casillas-Pérez, D., & Salcedo-Sanz, S. (2024). Two-step deep learning framework with error compensation technique for short-term, half-hourly electricity price forecasting. *Applied Energy*, *353*, 122059. https://doi.org/10.1016/j.apenergy.2023.122059

González-Romera, E., Jaramillo-Morán, M. A., & Carmona-Fernández, D. (2008). (F) Monthly electric energy demand forecasting with neural networks and Fourier series. *Energy Conversion and Management*, *49*(11), 3135–3142. https://doi.org/10.1016/j.enconman.2008.06.004

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Hajigholam Saryazdi, A. (2024a). *A Review of Global Electricity Market* (SSRN Scholarly Paper 4733363). https://doi.org/10.2139/ssrn.4733363

Hajigholam Saryazdi, A. (2024b). *Financial Dynamics of the Electricity Market* (SSRN Scholarly Paper 4855421). https://doi.org/10.2139/ssrn.4855421

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. Neural Computation. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, C.-J., Shen, Y., Chen, Y.-H., & Chen, H.-C. (2021). & A novel hybrid deep neural network model for short-term electricity price forecasting. *International Journal of Energy Research*, *45*(2), 2511–2532. https://doi.org/10.1002/er.5945

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *454*(1971), 903–995. https://doi.org/10.1098/rspa.1998.0193

Jiang, P., Nie, Y., Wang, J., & Huang, X. (2023). * Multivariable short-term electricity price forecasting using artificial intelligence and multi-input multi-output scheme. *Energy Economics*, *117*, 106471. https://doi.org/10.1016/j.eneco.2022.106471

Karim, M., Foysal, M., & Das, S. (2022). *Stock Price Prediction Using Bi-LSTM and GRU-Based Hybrid Deep Learning Approach* (p. 711). https://doi.org/10.1007/978-981-19-3148-2_60

Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, *221*, 386–405. https://doi.org/10.1016/j.apenergy.2018.02.069

Lago, J., De Ridder, F., Vrancx, P., & De Schutter, B. (2018). * Forecasting day-ahead electricity prices in Europe: The importance of considering market integration. *Applied Energy*, *211*, 890–903. https://doi.org/10.1016/j.apenergy.2017.11.098

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, *293*, 116983. https://doi.org/10.1016/j.apenergy.2021.116983

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. Proceedings of the IEEE. https://doi.org/10.1109/5.726791

Lehna, M., Scheller, F., & Herwartz, H. (2022). Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics*, *106*, 105742. https://doi.org/10.1016/j.eneco.2021.105742

Li, Q., Cheng, R., & Ge, H. (2023). Short-term vehicle speed prediction based on BiLSTM-GRU model considering driver heterogeneity. *Physica A: Statistical Mechanics and Its Applications*, *610*, 128410. https://doi.org/10.1016/j.physa.2022.128410

Li, X., Wang, H., Xiu, P., Zhou, X., & Meng, F. (2022). Resource Usage Prediction Based on BILSTM-GRU Combination Model. *2022 IEEE International Conference on Joint Cloud Computing (JCC)*, 9–16. https://doi.org/10.1109/JCC56315.2022.00009

Liu, H., & Chen, C. (2019). *Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy*, *249*, 392–408. https://doi.org/10.1016/j.apenergy.2019.04.188

Liu, H., Tian, H., Pan, D., & Li, Y. (2013). + Forecasting models for wind speed using wavelet, wavelet packet, time series and Artificial Neural Networks. *Applied Energy*, *107*, 191–208. https://doi.org/10.1016/j.apenergy.2013.02.002

Lu, H., Ma, X., Ma, M., & Zhu, S. (2021). # Energy price prediction using data-driven models: A decade review. *Computer Science Review*, *39*, 100356. https://doi.org/10.1016/j.cosrev.2020.100356

Meng, A., Wang, P., Zhai, G., Zeng, C., Chen, S., Yang, X., & Yin, H. (2022). Electricity price forecasting with high penetration of renewable energy using attention-based LSTM network trained by crisscross optimization. *Energy*, *254*, 124212. https://doi.org/10.1016/j.energy.2022.124212

Michael, N. E., Bansal, R. C., Ismail, A. A. A., Elnady, A., & Hasan, S. (2024). * A cohesive structure of Bi-directional long-short-term memory (BiLSTM) -GRU for predicting hourly solar radiation. *Renewable Energy*, *222*, 119943. https://doi.org/10.1016/j.renene.2024.119943

Musbah, H., & El-Hawary, M. (2019). (F) SARIMA Model Forecasting of Short-Term Electrical Load Data Augmented by Fast Fourier Transform Seasonality Detection. *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 1–4. https://doi.org/10.1109/CCECE.2019.8861542

Olah, C. (2015). *Colah's blog*. https://colah.github.io/posts/2015-09-NN-Types-FP/

Peng, L., Liu, S., Liu, R., & Wang, L. (2018). Effective long short-term memory with differential evolution algorithm for electricity price prediction. *Energy*, *162*, 1301–1314. https://doi.org/10.1016/j.energy.2018.05.052

Qian, Z., Pei, Y., Zareipour, H., & Chen, N. (2019). * A review and discussion of decomposition-based hybrid models for wind energy forecasting applications. *Applied Energy*, *235*, 939–953. https://doi.org/10.1016/j.apenergy.2018.10.080

Qiao, W., & Yang, Z. (2020). Forecast the electricity price of U.S. using a wavelet transform-based hybrid model. *Energy*, *193*, 116704. https://doi.org/10.1016/j.energy.2019.116704

Rana, M., & Koprinska, I. (2016). + Forecasting electricity load with advanced wavelet neural networks. *Neurocomputing*, *182*, 118–132. https://doi.org/10.1016/j.neucom.2015.12.004

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. IEEE Transactions on Signal Processing. https://doi.org/10.1109/78.650093

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Storn, R. (1997). Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *DIFFERENTIAL EVOLUTION*.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

transpower.co.nz. (2024). *Our System Operator role | Transpower*. https://www.transpower.co.nz/system-operator/our-system-operator-role/our-system-operator-role

Tschora, L., Pierre, E., Plantevit, M., & Robardet, C. (2022). - Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, *313*, 118752. https://doi.org/10.1016/j.apenergy.2022.118752

Weron, R. (2014). # Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Xiong, X., & Qing, G. (2023). A hybrid day-ahead electricity price forecasting framework based on time series. *Energy*, *264*, 126099. https://doi.org/10.1016/j.energy.2022.126099

Yao, S. J., Song, Y. H., Zhang, L. Z., & Cheng, X. Y. (2000). (FW) Wavelet transform and neural networks for short-term electrical load forecasting. *Energy Conversion and Management*, *41*(18), 1975–1988. https://doi.org/10.1016/S0196-8904(00)00035-2

Yu, X., Zhang, X., & Qin, H. (2018). (F) A data-driven model based on Fourier transform and support vector regression for monthly reservoir inflow forecasting. *Journal of Hydro-Environment Research*, *18*, 12–24. https://doi.org/10.1016/j.jher.2017.10.005

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0