

Article

Enhanced Data Processing and Machine Learning Techniques for Energy Consumption Forecasting

Jihye Shin ^{1,†} , Hyeonjoon Moon ^{2,†} , Chang-Jae Chun ³ , Taeyong Sim ³ , Eunhee Kim ⁴  and Sujin Lee ^{3,*} ¹ Department of Artificial Intelligence, Sejong University, Seoul 05006, Republic of Korea; jyes945@gmail.com² Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; hmoon@sejong.ac.kr³ Department of Artificial Intelligence and Data Science, Sejong University, Seoul 05006, Republic of Korea; cchun@sejong.ac.kr (C.-J.C.); tysim@sejong.ac.kr (T.S.)⁴ Department of Defense System Engineering, Sejong University, Seoul 05006, Republic of Korea; eunheekim@sejong.ac.kr

* Correspondence: genegraphy@sejong.ac.kr

† These authors contributed equally to this work.

Abstract: Energy consumption plays a significant role in global warming. In order to achieve carbon neutrality and enhance energy efficiency through a stable energy supply, it is necessary to pursue the development of innovative architectures designed to optimize and analyze time series data. Therefore, this study presents a new architecture that highlights the critical role of preprocessing in improving predictive performance and demonstrates its scalability across various energy domains. The architecture, which discerns patterns indicative of time series characteristics, is founded on three core components: data preparation, process optimization methods, and prediction. The core of this architecture is the identification of patterns within the time series and the determination of optimal data processing techniques, with a strong emphasis on preprocessing methods. The experimental results for heat energy demonstrate the potential for data optimization to achieve performance gains, thereby confirming the critical role of preprocessing. This study also confirms that the proposed architecture consistently enhances predictive outcomes, irrespective of the model employed, through the evaluation of five distinct prediction models. Moreover, experiments extending to electric energy validate the architecture's scalability and efficacy in predicting various energy types using analogous input variables. Furthermore, this research employs explainable artificial intelligence to elucidate the determinants influencing energy prediction, thereby contributing to the management of low-carbon energy supply and demand.

Keywords: carbon neutrality; data preprocessing optimization; energy efficiency; energy supply and demand; machine learning; time series analysis



Citation: Shin, J.; Moon, H.; Chun, C.-J.; Sim, T.; Kim, E.; Lee, S. Enhanced Data Processing and Machine Learning Techniques for Energy Consumption Forecasting. *Electronics* **2024**, *13*, 3885. <https://doi.org/10.3390/electronics13193885>

Academic Editor: Ahmed Abu-Siada

Received: 20 August 2024

Revised: 23 September 2024

Accepted: 27 September 2024

Published: 30 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Climate change and global greenhouse gas emissions require an approach to global sustainable energy production and consumption [1]. In the pursuit of carbon neutrality and improved energy efficiency, a stable energy supply is necessary. To effectively manage high energy demand, it is important to accurately predict demand in advance and ensure energy supply stability [2]. Notably, within the context of carbon emission reduction and urban energy management, the prediction of heat supply in district heating system (DHS) has emerged as a significant concern [3]. Accurately predicting heat supply in DHS can contribute to energy consumption optimization, cost savings, efficient facility management, and energy supply stability [4].

Data collected over time, such as heat supply in DHS, typically take the form of time series. The diverse types, complexities, seasonal patterns, and temporal structures of time series data make actual predictions challenging [5]. At this time, structural issues, such as

the generation of incomplete data during the creation of extensive time series data, further compound the complexity of real-world predictions [6]. In other words, if time series data preprocessing methods are inaccurate or ineffective, it can lead to a degradation in the performance of predictive models.

A deep learning-based architecture is proposed that emphasizes preprocessing techniques to effectively harness time series data and address these challenges. Effective preprocessing methods are first focused on, accounting for the complex patterns and characteristics of time series data. The data are then optimized through various scenarios, considering the inherent characteristics of time series, to enhance the performance of predictive models. The generality of the proposed preprocessing methodology is validated by applying it to various machine learning (ML) algorithms and energy types. Performance comparisons and analyses of multiple ML models using the optimized data are conducted. Furthermore, the same methodology is applied to both thermal and electrical energy, confirming its applicability to various energy types. In the final analysis, explainable artificial intelligence (XAI) is employed to identify variables that affect energy prediction and to elucidate their contributions. Using the XAI methodology, the primary factors influencing energy output were identified and analyzed.

The main contributions of this study are as follows. First, the proposed architecture emphasizes the importance of preprocessing and improves predictive performance by considering various factors in time series prediction. Second, performance improvement was confirmed through comparative analysis between various models using optimized data. Third, the proposed architecture shows the potential for expansion into other types of energy prediction areas as well as heat energy.

Precise predictions of heat supply within DHS are essential for optimizing energy consumption, reducing expenses, managing facilities efficiently, and guaranteeing the stability of energy supply. It is anticipated that this study will significantly enhance energy prediction capabilities, thereby contributing to greater energy efficiency and cost reduction.

The structure of this paper is as follows: Section 2 reviews the literature on the increasing importance of efficient energy utilization and advancements in energy prediction research. Section 3 provides an overview of the proposed architecture for energy prediction, with a focus on time series data processing. Section 4 presents empirical results, summarizing the data used in experiments with heat and electric energy sources. Finally, Section 5 concludes the paper by summarizing the proposed time series prediction architecture and providing insights into its implications and future work.

2. Literature Review

As the necessity for effective energy utilization continues to grow, the field of energy consumption forecasting has become a pivotal area of investigation. A wide range of ML models has been investigated for their potential to predict energy usage across various systems, particularly in DHS, which plays a pivotal role in providing heat to residential and commercial buildings [7]. The most prominent models used for these tasks include extreme gradient boosting (XGBoost) [8], light gradient-boosting machine (LightGBM) [9], categorical boosting (CatBoost) [10], multilayer perceptron (MLP) [11], and long short-term memory (LSTM) networks [12], which are particularly adept at handling time series data and capturing complex patterns. Furthermore, the utilization of preprocessing techniques is of paramount importance in order to optimize the accuracy of these models [13]. This is due to the fact that such techniques guarantee the quality and relevance of the input data, which is of critical importance for the generation of reliable predictions. XGBoost and LightGBM are gradient-boosting models that have been extensively employed in the field of energy consumption forecasting, particularly in the context of district heating systems. These models are renowned for their capacity to process voluminous datasets in an expedient and parsimonious manner, while simultaneously mitigating the phenomenon of overfitting [14].

XGBoost, which employs a tree-based boosting approach, is capable of handling missing data and noisy inputs. It has been successfully employed in heat load forecasting due to its robustness and capacity to discern nonlinear relationships in data [15,16]. LightGBM, which employs a leaf-wise tree growth strategy, has demonstrated superior efficiency in terms of speed and memory usage compared with XGBoost [17,18]. This renders it particularly suitable for real-time heat load prediction in district heating systems. LightGBM is capable of handling both continuous and categorical variables, thereby making it a versatile choice for energy consumption forecasting. However, both models are contingent upon the quality of the data preprocessing, including feature scaling, normalization, and missing value imputation. CatBoost, another gradient boosting model, has been optimized for the handling of categorical variables without the necessity for extensive preprocessing, such as one-hot encoding [14]. In the context of district heating energy forecasting, CatBoost is particularly advantageous as it reduces the preprocessing time and effort while maintaining high accuracy in predictions. It has been successfully applied to energy consumption prediction tasks where data types vary significantly, allowing for the seamless integration of both numerical and categorical data. MLP models, which are neural network-based, are effective at modeling nonlinear relationships in energy consumption data.

However, it should be noted that MLPs are sensitive to the scale and distribution of input data, which makes the implementation of preprocessing steps such as feature scaling and normalization a critical aspect for achieving high performance. It is essential to undertake comprehensive preprocessing of MLPs to avoid overfitting, particularly when working with complex and noisy datasets that are prevalent in district heating systems. LSTM networks are a specific type of recurrent neural network that has been designed to capture temporal dependencies in time series data. This makes them an especially suitable choice for energy consumption forecasting in district heating systems [15,19]. LSTMs are capable of modeling long-term dependencies, such as seasonal variations in heat demand [20]. Nevertheless, LSTMs necessitate comprehensive preprocessing, encompassing sequence padding, normalization, and the handling of missing values to achieve optimal performance. The complexity of the data and the necessity for meticulous feature selection serve to underscore the significance of preprocessing.

The efficacy of these machine learning models is contingent upon the quality of the preprocessing techniques employed. The application of data preprocessing techniques, including data cleaning, imputation, normalization, and feature selection, is essential for enhancing the accuracy of energy forecasting predictions. The significance of data analysis in enhancing the efficacy of predictive models has become apparent, such as in the studies referenced in [21,22]. Additionally, there has been a notable increase in research studies that focus on model interpretation with XAI. One study utilizes XAI to analyze the relationship between energy consumption and input variables, revealing that energy prediction outcomes may vary depending on factors such as the season, indoor and outdoor conditions, and operational methods [23]. Other researchers employ XAI technology to interpret the results and assess the importance of variables used in predictions [24,25]. Additionally, another study proposes a methodology for selecting input variables for energy consumption prediction based on the variable importance derived from XAI [26]. In district heating systems, where energy consumption patterns are influenced by various external factors such as weather conditions, preprocessing is beneficial in reducing noise, handling missing data, and ensuring that the models focus on the most relevant features. Furthermore, the prediction of building energy consumption has also been the subject of considerable research, particularly in the context of sustainable urban development. It is evident that residential and commercial buildings account for a considerable proportion of global energy consumption. Therefore, it is imperative to accurately forecast energy usage in these buildings in order to maintain equilibrium between supply and demand [14,27]. A variety of machine learning models, including gradient boosting decision trees (GBDTs), support vector machines (SVM)s, and hybrid models, have been extensively utilized for the purpose of predicting energy consumption in buildings. These models utilize historical

energy data, meteorological information, and building-specific parameters to forecast future energy demand, thereby optimizing energy distribution and reducing inefficiencies [7].

The key preprocessing techniques employed include the imputation of missing data, the normalization of variables, and the scaling of features, as well as the selection of pertinent features. This ensures that models such as XGBoost and LightGBM are able to handle incomplete datasets without any loss of prediction accuracy. It is essential to ensure consistent training to prevent the dominance of features with larger ranges for models, as well as identify the most important variables influencing heat load, thereby reducing the complexity of the model and improving prediction performance.

Finally, XAI techniques are being increasingly applied in energy forecasting to enhance the interpretability of machine learning models. XAI helps in identifying the most critical factors that influence energy predictions, thereby improving both the trustworthiness and reliability of the models. In previous studies, individual proposals were focused on energy prediction, data preprocessing, data analysis, and the utilization of XAI separately. Accordingly, this study proposes an integrated energy prediction architecture that encompasses data preprocessing, analysis, and XAI, addressing all these aspects collectively.

3. Architecture Overview

This section provides an overview of the proposed architecture for energy prediction, focusing on time series data processing. As depicted in Figure 1, the architecture comprises three main stages: data preparation, process optimization methods, and prediction. The data preparation phase collects data based on objectives and synchronizes the collected data in time. In the process optimization methods stage, data are optimized based on three conditions that consider the patterns and characteristics of time series. These conditions include normalization, data cleaning, data split patterns, and data split ratio to enhance model performance, reduce training time, and mitigate the impact of outliers. Min-max normalization is applied to address the varying range of feature values, ensuring consistent data scaling. Additionally, eight data cleaning methods are considered to resolve issues such as missing values and duplication. The data split patterns are designed to reflect seasonal and quarterly trends inherent in time series data, while the data split ratios are set to capture the overall trends over time, ensuring a robust and accurate model. Subsequently, the optimized data are used to compare and analyze the performance of prediction models.

Model optimization was performed using various methods to enhance performance. In this study, the data are optimized through enhanced preprocessing and cleaning of time-series energy data, along with feature engineering using Shapley additive explanations (SHAP) [28] to achieve optimal performance. Feature engineering transforms input data by selecting important features or creating new ones, thereby improving the model's predictive capabilities. Data preprocessing and cleaning enhance data quality by addressing missing values, removing outliers, and standardizing the dataset. Cross-validation and data splitting ensure proper division of the data into training, validation, and test sets, facilitating robust evaluation of the model's generalizability and deriving optimal results.

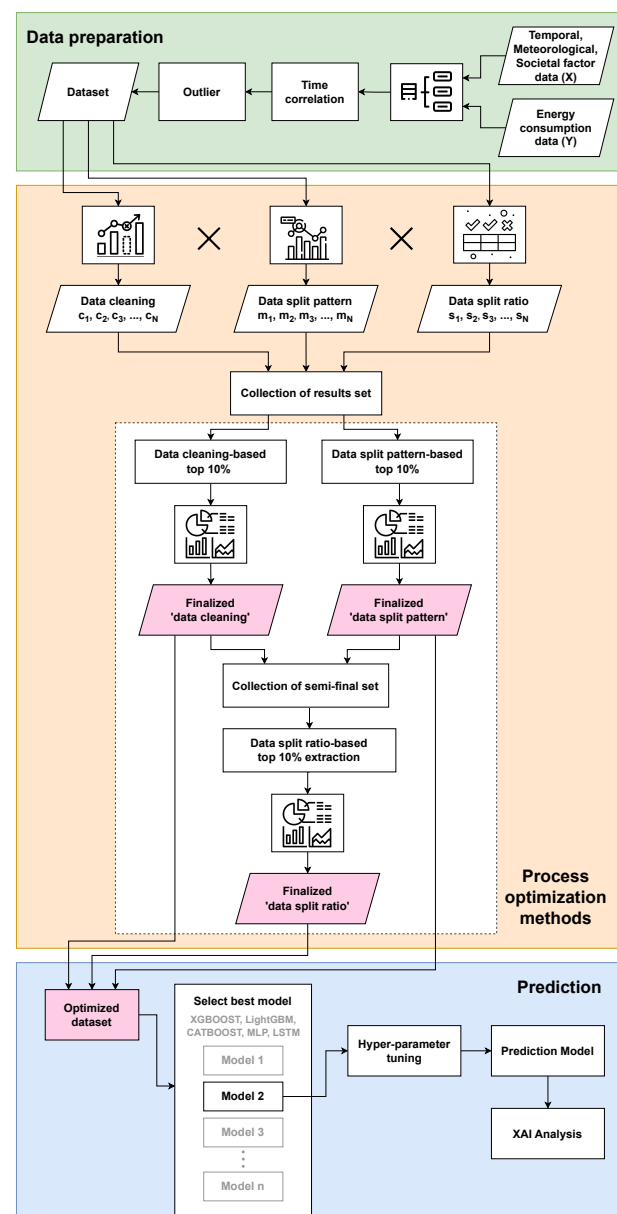


Figure 1. Architecture for energy prediction.

3.1. Data Preparation

This section describes the data preparation process in terms of data collection and time correlation. First, it discusses the data collection for the experiment. Subsequently, it examines time correlation to analyze the impact of time on the data.

3.1.1. Data Collection

Data collection is determined based on the objectives of the predictive model. For energy prediction, energy usage is used as the target variable choosing the input variables accordingly. The significant factors influencing energy consumption can mainly be categorized into three groups: meteorological factors, temporal factors, and societal factors [29]. From these three categories, variables commonly used in energy prediction were selected.

Table 1 presents the input variables used for energy prediction, selected based on their relevance to energy consumption. Meteorological data use data from the same period in which energy data are collected and represent synoptic-scale weather data for the area where energy data are utilized. Synoptic meteorological observations involve collecting

weather data simultaneously at multiple observation stations at specified times. The collected variables include temperature, wind speed, wind direction, humidity, dew-point temperature, local atmospheric pressure, sunshine duration, solar radiation, visibility, and ground temperature. Additionally, time factors such as year, month, day, and hour are utilized as variables based on the collected time information. Period-specific issues, such as the influence of extended periods like COVID-19 and holidays identified in time data, are considered social factors. The collected data are prepared to align with the temporal flow, combining data from the same periods into a single dataset. To merge into one dataset, the time intervals of the data need to be synchronized.

Table 1. Input variables used in the experiment.

Data Type	Input Variable
Temporal factor	year, month, day, hour
Meteorological factor	temperature, wind speed, wind direction, humidity, dew-point temperature, local atmospheric pressure, sunshine duration, solar radiation, visibility, ground temperature
Societal factor	issues by period

3.1.2. Time Correlation

Since time plays an important role in time series prediction, it is important to set the data sampling interval and the time interval between the input and output variables of the model. In addition, this setting should be adapted for each purpose and can have a great influence on the accuracy and efficiency of the prediction model.

Two different methods are applied to set the time points for input and output variables based on the purpose of the prediction day. Figure 2 illustrates the two approaches for setting the time points of input and output variables. The first approach involves setting the time points of input and output variables to be the same. In Figure 2a, as shown, the approach involves training pairs of input and output variables at the same time period. This method utilizes predicted climate information and other data for the desired day as inputs, such as time and holidays, to forecast the future. The second approach involves setting the time points of input and output variables differently. As shown in Figure 2b, this method pairs current-time input variables with future-time output variables for training. In this approach, the perspective can also vary depending on the time interval between the present and the future. For input variables $\mathbb{X} = \{x_1, x_2, x_3, \dots, x_n\}$ and output variables $\mathbb{Y} = \{y_1, y_2, y_3, \dots, y_n\}$, approach (a) computes as $y_i = x_i$, and approach (b) becomes $y_i = x_{i+t}$, where t is the time interval between input variables and output variables. In this paper, data are collected by time for energy prediction, and the first approach is used as a method for time correlation.

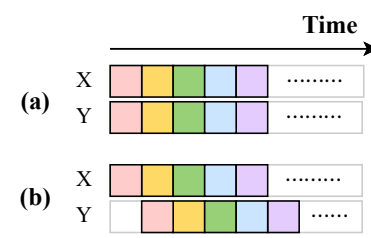


Figure 2. Approaches for time correlation: (a) Same time points for input and output variables. (b) Different time points for input and output variables.

3.2. Process Optimization Methods

In AI-driven time series prediction models, data hold immense significance. Data assist in comprehending the relationships between the input variables of the model and the output variables targeted for prediction. Consequently, a plethora of factors, including correlations,

temporal lag relationships, seasonality, and more, must be considered. Analyzing the data to uncover these relationships and incorporating them into the model can enhance prediction accuracy [30]. As the accuracy of the model's predictions is heavily dependent on the quality and quantity of the data, creating an AI prediction model that is insensitive to data is generally unattainable. In developing AI prediction models, especially for time series data, it is essential to prioritize data sensitivity. Time series data are intrinsically linked to time progression, with their unique patterns significantly impacting prediction accuracy. For this experiment, the commonly used LightGBM, known for its faster speed compared with other regression models, is employed in time series prediction.

Dataset normalization is a crucial technique in ML, especially when features have varying ranges [31–33]. This process can enhance model performance and reduce training times [31,33]. Among numerous normalization methods, min–max normalization is widely validated and used in the literature due to its effectiveness in preserving the relationships between data points and its simplicity in implementation [34–40]. The MinMaxScaler ensures that feature values are scaled to a range between 0 and 1, which mitigates the impact of outliers and reduces standard deviation [41,42]. This bounded range is particularly advantageous in algorithms sensitive to the magnitude of input data, such as neural networks and distance-based methods [43,44]. The MinMaxScaler function from the Scikit-Learn library [45] was employed to normalize the dataset features within the 0 to 1 range during both the training and testing phases of the model, following recommendations and findings from recent studies. The main considerations include three factors: data cleaning, data split patterns, and data split ratio.

3.2.1. Data Cleaning

The first condition considered in the experiment is data cleaning. Time-series data inevitably encounter issues such as missing values and duplicates during the collection process [46]. Data cleaning is essential to solve these problems. Data cleaning is the process of discovering and correcting damaged or inaccurate records and involves identifying, replacing, correcting, or deleting incomplete or inaccurate parts of data. An appropriate method is selected by considering eight missing value processing techniques. Commonly used methods for handling missing values in time series data are deletion, imputation, and prediction [47]. Deletion involves removing samples or variables with missing values, imputation replaces missing values with other values, and prediction entails estimating missing values.

In this experiment, for handling missing values, eight methods are employed based on the treatment approach: one deletion method and seven imputation methods, resulting in eight datasets. Each method employs a widely utilized approach for processing missing data in time series prediction [48–51].

The first approach (dataset0) is the deletion of missing values. Removing missing values can enhance predictive performance compared with using improperly refined missing values. The second approach (dataset1) involves replacing missing values with the preceding data. In essence, missing values are replaced with the preceding data, and if those are missing too, with the data before them. Chen et al. [52] used this approach to replace missing data in time series datasets with highly correlated data in terms of attributes and time. The third approach (dataset2) replaces missing values with data from the subsequent hour. Missing values are replaced with subsequent data, and if those are missing too, with data further ahead. The fourth approach (dataset3) employs linear interpolation to replace missing values. Linear interpolation approximates a function that fits the given data points and calculates missing values for the variable using this equation. This method effectively compensates for continuously missing data while considering the dynamic nature of the dataset [53,54]. The fifth approach (dataset4) replaces missing values with the moving average. The moving average period is set as one day (24 h). The method involves calculating the average of nonmissing values within the chosen period and using that calculated value to replace missing values within that period. The process is iterated

by shifting the period by one interval. Menezes et al. [55] implemented a simple moving average approach for extracting trend, seasonal, and residual components. The sixth approach (dataset5) replaces missing values with the moving median, which is similar to the moving average, but median values are used instead. The seventh approach (dataset6) replaces missing values with the overall mean. The overall mean is calculated from all available data from 1 January 2012 to 31 December 2021. The eighth approach (dataset7) replaces missing values with the overall median, similar to dataset6.

Among the eight preprocessing methods described, the most suitable method was selected based on its ability to enhance predictive performance while maintaining data integrity.

3.2.2. Data Split Patterns

The second consideration is data split patterns. Time series data are generally collected and utilized in chronological order. In the case of seasonal time series data, different patterns may appear for each specific quarter. Data are patterned and divided in consideration of the characteristics of such time series data.

Figure 3 illustrates the concept of data splitting using months as dividing points, explaining the experimentation involving divisions into 12 months, 6 months, 4 months, and 1 month for a span of 2 years. In Figure 3a, for instance, the division into 12 months involves constructing a model that considers the entire temporal flow from January to December. In this method, a total of 1 model is considered. In Figure 3b, the division into 6 months separates the data into two halves: the first half (January to June) and the second half (July to December) to create two separate models. Here, two models are considered, corresponding to the first and second halves of the year. In Figure 3c, dividing into 4 months involves creating models based on three quarters, resulting in three models being considered. Lastly, in Figure 3d, the division into 1 month means constructing models for each individual month. This method results in a total of 12 models, each corresponding to a specific month. In the figure, data with the same color are trained together in a single model. Considering data split patterns in the above way helps avoid being bound by the seasonal characteristics of the data.



Figure 3. An example of data split patterns: (a) Division into 12 months: a single model. (b) Division into 6 months: two models. (c) Division into 4 months: three models. (d) Division into 1 month: twelve models.

3.2.3. Data Split Ratio

The third consideration is the data split ratio. In this study, training and test data were divided into one-year increments for annual energy prediction. The reason for aiming at annual predictions is to assess whether the model can effectively predict changes and capture the overall trends in the flow of time series data. While this provides a comprehensive overview of seasonal variations, predicting beyond one year becomes less reliable due to recent extreme weather events and the shortening of seasonal cycles. These factors introduce significant variability, making predictions over longer periods, such as

three or ten years, less efficient and more prone to inaccuracies. Figure 4 illustrates how the performance of the model trained on the heat energy dataset from 2012 to 2016 changed over time. To evaluate performance, the R-squared score (R2 score), mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) were used. Detailed information about these metrics can be found in Section 3.3. The figure shows a consistently decreasing trend in all performance metrics, except for 2019. Additionally, it confirms that the performance after 2 years does not exceed the performance immediately after training.

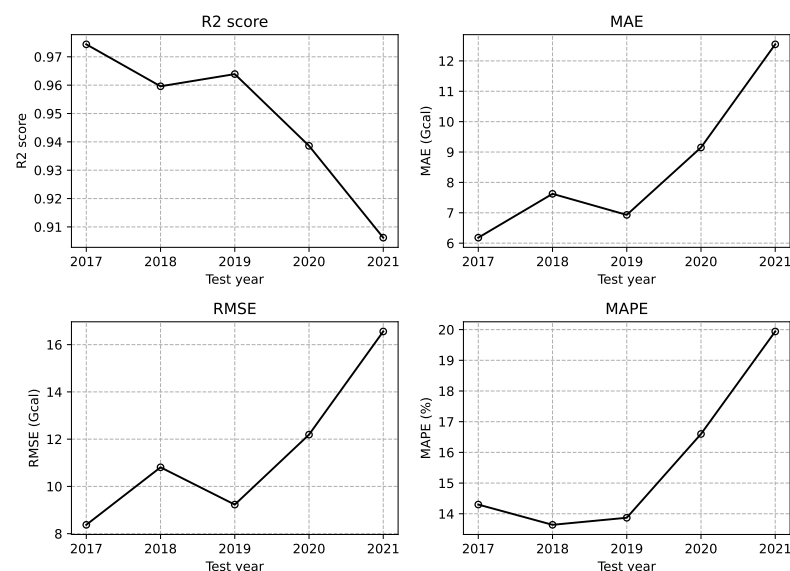


Figure 4. Model performance over time: heat energy data (training: 2012–2016).

Accordingly, one year of data is designated as test data. And the percentage of the data split ratio is determined by the number of years of training data utilized. For instance, when training a model using one year's worth of data and testing it with another year's worth of data, the ratio of training to testing data is set at 50:50. Similarly, if you train the model with three years of data and use one year's data for testing, the training-to-testing data ratio is adjusted to 75:25. By varying the amount of data and the training-to-testing data ratio in this manner, one can consider the impact of data quantity on model performance.

3.2.4. Approach to Searching for the Best Condition

In this study, experiments are conducted to prepare optimized data based on the three mentioned considerations. After the experiments are conducted, the results are analyzed to finalize the optimal conditions for each consideration.

Figure 5 illustrates the methodology for identifying the optimal conditions. Initially, the three conditions, "Data cleaning", "Data split pattern", and "Data split ratio", are combined to create "Categorized cases". The results of the "Categorized cases" are then sorted by highest performance for each metric and combined by selecting the top 10%. In this case, "Combined cases" represent 40% of the total data and may include duplicate values. In the "Combined cases", the condition value with the highest percentage for the condition of interest is identified and selected. In this process, the values for "Data cleaning" and "Data split pattern" are determined, and for the results where these values are used, each metric is sorted in order of highest performance, the top 10% are selected, and they are combined. Additionally, the cases where the "data split ratio" is used the most are identified. Using the selected conditions, a final dataset for model training is prepared. This process allows for the identification of the most appropriate conditions for optimizing the dataset, which can subsequently be used to train the final model.

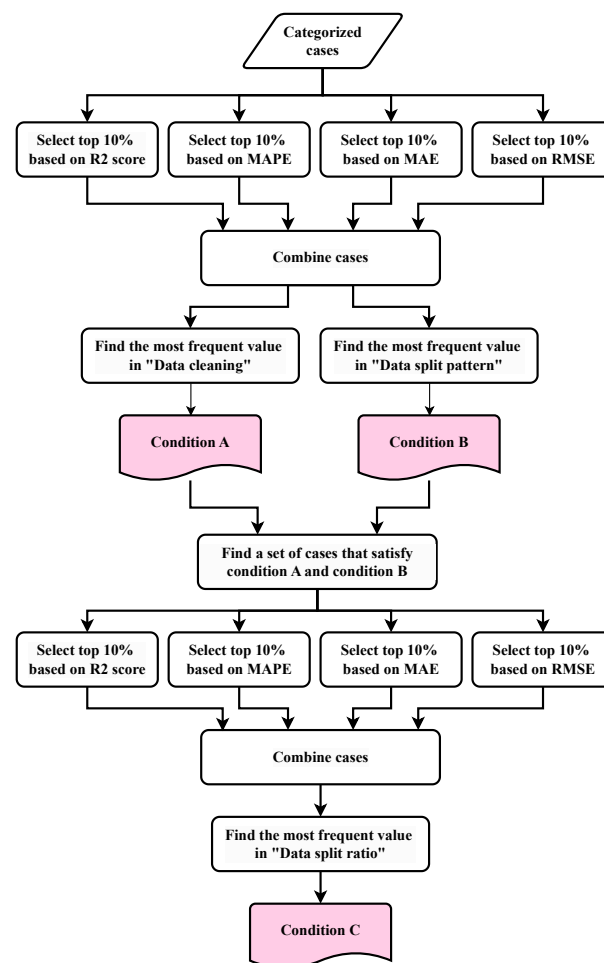


Figure 5. A flowchart of an approach to searching for the best condition.

3.3. Prediction

In this experiment, five models—XGBoost, LightGBM, CatBoost, MLP, and LSTM—were adopted to validate the effectiveness of the proposed method on time series energy data. Each of these models offers unique strengths that align with the characteristics of the dataset [8–12,56–62]. XGBoost [8] and LightGBM [9] are particularly effective for data due to their exceptional ability to handle large datasets with high dimensionality. Their gradient-boosting framework captures complex patterns and interactions within time series data, leading to significant improvements in predictive accuracy. CatBoost [10] is proficient in handling categorical features inherent in datasets, ensuring that important categorical variables are efficiently incorporated into the model, thereby enhancing predictive performance. MLP [11] is capable of modeling nonlinear relationships, making it suitable for capturing the intricate temporal dependencies in time series data. Its ability to learn from complex, nonlinear patterns contributes to its effectiveness in application. LSTM [12] excels in processing sequential data and retaining long-term dependencies, making it particularly suitable for time series data where capturing trends over extended periods is crucial for accurate predictions. These models have demonstrated robustness and effectiveness in handling specific time series energy data, leading to substantial improvements in performance. By leveraging these proven techniques, comprehensive and reliable predictions were ensured in this study.

The performance of these models is evaluated using four metrics with different strengths and weaknesses to ensure the robustness of predictions. R2 score, MAE, RMSE, MAPE, and training time are used to evaluate comprehensive performance. The R2 score [63,64] is a metric that measures how well a model explains the variance in the

dependent variable, ranging between 0 and 1. A value close to 1 signifies that the model effectively explains the data. R2 score evaluates the goodness of fit of the model, making it useful for assessing suitability in the context of energy prediction. MAE [65] represents the average of the absolute errors between predicted values and actual values, indicating the size of errors. It measures the size of the model's prediction errors, making it suitable for assessing how accurate the model's predictions are and understanding the model's prediction quality. RMSE [65,66] is a metric that represents the square root of the MSE, which measures the average of the squared errors between predicted values and actual values. As it uses squared errors to assess the predictive quality of a model, it is suitable for understanding the quality of prediction errors intuitively. MAPE [67] is a metric that represents the average of the absolute percentage errors between predicted values and actual values. It assesses relative prediction quality by considering errors expressed as percentages.

Five predictive models and four evaluation metrics are used to check their performance on optimization from a data perspective. And a model demonstrating evenly high performance across the four evaluation indicators is selected. Finally, the application of XAI is discussed to enhance reliability by showing the analyzed patterns and which input variables influenced the AI model's results. XAI helps determine the importance of variables by revealing how input variables influence the model's results [68]. SHAP [28], one of the XAI models, is used in this study. SHAP represents the impact of features used in a model on the results by utilizing Shapley values, a measure rooted in the concept of game theory for determining feature importance. It visualizes how each feature influences results based on actual predictions and mean predictions, fostering user confidence in the outcomes of AI models.

4. Empirical Results

This section outlines the data utilized in the experiments, focusing on heat and electric energy sources, and details the experimental process and the results of applying the proposed architecture to predict energy consumption patterns. Experiments were conducted on the proposed architecture using district heat energy for energy prediction. To verify its scalability to various domains, experiments are conducted not only with district heat energy but also with electric energy.

4.1. Heat Energy in Cheongju, Republic of Korea

This section presents the experiments conducted using the proposed architecture with heat energy data. Section 4.1.1 provides a description of the data utilized in the experiments. Section 4.1.2 outlines the process optimization methods, as described in Section 3.2. Finally, Section 4.1.3 compares and evaluates five different models based on the results obtained in Section 4.1.2 across various scenarios.

4.1.1. Dataset Description

Heat supply data, used to predict actual heat energy consumption, were collected hourly from an eco-friendly liquefied natural gas combined heat and power plant in Cheongju, Republic of Korea. The data are collected on an hourly basis from 2012 to 2021, comprising a total of 87,672 data points. Heat supply refers to the amount of heat delivered from the power plant to consumers, and the unit is gigacalories (Gcal). It covers heat consumption for space heating and domestic hot water usage. Figure 6 visualizes the profile of the heat energy usage, showing a recurring pattern with a one-year cycle, reflecting its seasonal characteristics.

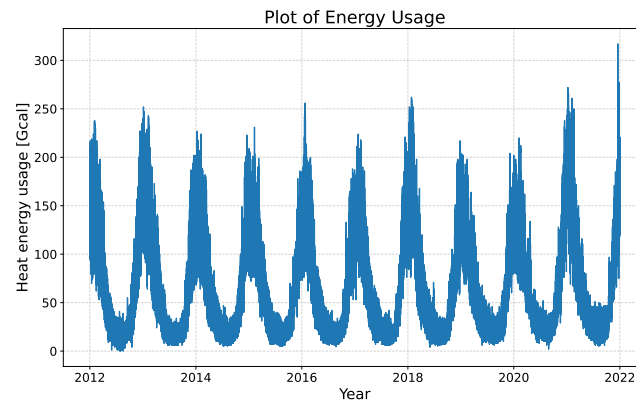


Figure 6. A plot of heat energy usage.

To understand the periodicity of energy usage and its correlation with the outdoor temperature, one of the key variables, the monthly average energy usage and monthly average temperature are visualized in Figure 7. The x -axis represents the months, and the y -axis represents the monthly average energy usage and the monthly average temperature. The monthly average energy and temperature are computed by summing respective values across different years for each month and dividing by the corresponding number of data points. This visualization reveals that energy usage is high in the low-temperature period and low in the high-temperature period, indicating a seasonal pattern in energy consumption.

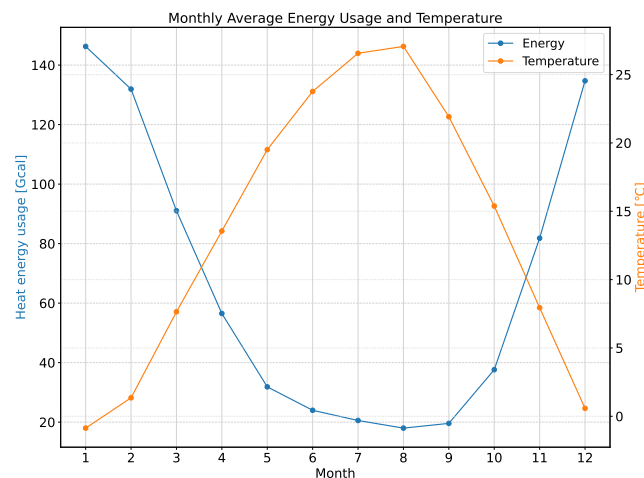


Figure 7. Monthly average heat energy usage and temperature.

Figure 8 showed the relationship between daily mean temperature and energy usage deviation. It shows that higher temperatures generally lead to a decrease in energy usage deviation, whereas lower temperatures cause an increase.

Heat energy demand is influenced by climate-related meteorological variables, including outdoor temperature, wind speed, solar radiation, humidity, and precipitation [69]. The weather data consist of hourly weather data for Cheongju and are used for predicting hourly heat supply in Cheongju. The latitude and longitude coordinates for Cheongju are approximately 37.5714, 126.9658. The minimum, maximum, mean, and standard deviation of the meteorological variables and heat energy used for prediction are described in Table 2.

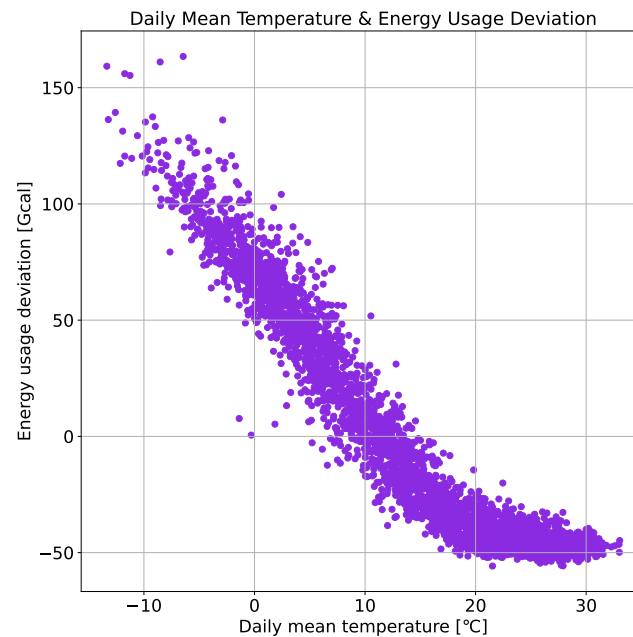


Figure 8. Daily mean temperature and heat energy usage deviation.

Table 2. The minimum, maximum, mean, and standard deviation of meteorological variables and heat energy for Cheongju.

Name	Minimum	Maximum	Mean	Standard	Unit
Datetime	1 January 2012	31 December 2021	-	-	-
Temperature	−16.5	38.1	13.76	10.84	°C
Wind speed	0	8.7	1.47	0.94	m/s
Wind direction	0	360	200.55	114	16 directions
Humidity	7	100	61.32	20.03	%
Dew-point temperature	−28.1	26.3	5.6	11.49	°C
Local atmospheric pressure	975.1	1032.1	1009.6	8.19	hPa
Sunshine duration	0	1	0.5	0.45	h
Solar radiation	0	3.86	1.06	0.94	MJ/m ²
Visibility	0	6454	1597.799	46.93	10 m
Ground temperature	−10.9	62.2	15.12	12.2	°C
Energy	0	317	65.9	52.92	Gcal

4.1.2. Process Optimization Methods

The dataset of heat energy was optimized using the defined process optimization methods described in Section 3.2. In the process optimization methods, a total of 1440 scenarios are generated based on the conditions. Under the proposed architecture, 1440 scenarios were generated using eight data cleaning methods, four data split patterns, and 45 different combinations of data split ratios and years. The data split ratio of 1440 scenarios is shown in Table 3.

First, the data split pattern group selects data cleaning candidates for condition1 optimization that are the most frequently used. Looking at the percentage occupied by each evaluation metric, R2 score is mostly occupied by dataset1, dataset2, and dataset3, among which dataset2 and dataset1 show higher performance. For MAE and RMSE, dataset0, dataset1, and dataset3 occupy the highest percentage, with dataset0 and dataset1 showing better performance. For MAPE, dataset1, dataset4, and dataset5 occupy the highest percentage, with dataset1 showing the best performance. Overall, dataset1 appears to be the optimized data cleaning method.

Similarly, for data split ratios, the distribution was examined, and the most frequently used data split ratio methods were selected as candidates for condition3 optimization.

Looking at the percentage occupied by each evaluation metric, R2 score is mostly occupied by 4 months and 6 months, with 4 months showing better performance. For MAE, 6 months and 4 months occupy the highest percentage, with 6 months showing better performance. For RMSE, 12 months and 6 months occupy the highest percentage, with 12 months showing better performance. For MAPE, 6 months and 12 months occupy the highest percentage, with 12 months showing better performance. Overall, 12 months appears to be the optimized data split pattern method.

Table 3. The train-to-test ratio of training data and test data for heat energy prediction.

Train Ratio	Test Ratio	Count	Percentage of Count [%]
50	50	288	20.00
67	33	256	17.78
75	25	224	15.56
80	20	192	13.33
83	17	160	11.11
86	14	128	8.89
88	13	96	6.67
89	11	64	4.44
90	10	32	2.22
Total		1440	100

Finally, to select the data split ratio, the same data cleaning methods and data split patterns are used. Data cleaning and data split patterns are grouped for dataset1 and 12 months, which were previously selected using the optimized method. The top 10% based on performance metrics are sorted. Based on the sorted top data, the data split ratio that shows the highest performance is determined. Despite the differential training of data split ratios, it can be confirmed that the case of training with a ratio of 83:17 shows the highest performance. The experiment method selected in this manner shows an improved performance compared with training with all data except the test data.

Table 4 presents the maximum, minimum, and average values for each metric in the “Categorized cases” category. This category encompasses 1440 cases, which were obtained by combining all eight data cleaning methods, four data split pattern methods, and 45 methods that consider data split ratio and frequency for heat energy. The results demonstrate that the R2 score, MAE, RMSE, and MAPE may vary by up to 0.2001, 7.3107 Gcal, 15.7568 Gcal, and 6.1940e+12%, respectively, depending on the case selected. It can be observed that for 0.9760, the maximum values of the R2 score are observed in all cases, with MAE, RMSE, and MAPE values of 6.1373 Gcal, 8.3646 Gcal, and 12.85%, respectively. These values do not demonstrate the optimal performance against other indicators. Similarly, when MAE is at its optimal performance (5.7721 Gcal), R2 score, RMSE, and MAPE are 0.9666, 7.9720 Gcal, and 13.36%, respectively. When RMSE is at its optimal performance (7.9365 Gcal), R2 score, MAE, and MAPE are 0.9642, 5.8812 Gcal, and 14.34%, respectively. When MAPE is at its optimal performance (11.93%), R2 score, MAE, and RMSE are 0.9718, 5.9272 Gcal, and 8.1546 Gcal, respectively. Consequently, it is crucial to consider each indicator equally, as the performance of the model may vary depending on which indicator is prioritized. Accordingly, in the proposed method, the data cleaning method and the data split pattern method are selected through comparative analysis by selecting the top 10% for each evaluation index. The performance of each evaluation index for the conditions that satisfy the selected methods can be found in Table 5. It should be noted that the highest performance value observed in Table 5 may not correspond to that observed in Table 4. However, a comparison of the average values reveals that all of them demonstrate an improvement in performance.

Table 6 presents the results of training the LightGBM model on heat energy using two different time periods: 5 years and 10 years. It shows performance improvements with

an R2 score of 0.0019, MAE of 0.1612, RMSE of 0.0882, and MAPE of 0.56, confirming that using a large amount of data does not necessarily result in better performance immediately.

Table 4. Categorized cases performance summary for heat energy dataset.

Metrics	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]
Max	0.9760	13.0828	23.6933	27.93
Mean	0.9450	7.7335	11.3172	15.21
Min	0.7759	5.7721	7.9365	11.93
Count	1440	1440	1440	1440

Table 5. A set of cases performance summary for heat energy dataset which satisfied selected data cleaning and data split pattern.

Metrics	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]
Max	0.9759	9.9506	14.3417	19.33
Mean	0.9613	7.1773	10.0046	14.10
Min	0.9296	6.0865	8.3587	12.07
Count	45	45	45	45

Table 6. Comparison of LightGBM training results using 5 years of heat energy data and 10 years of heat energy data.

Training Duration	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]
10 years	0.9572	7.9886	11.1824	14.18
5 years	0.9591	7.8274	10.9330	13.62

4.1.3. Prediction

After data optimization, the performance of the five models is compared in different scenarios. As shown in Table 7, each scenario was trained with different periods, and other conditions were based on the process optimization method.

Table 7. The period of each scenario for heat energy prediction.

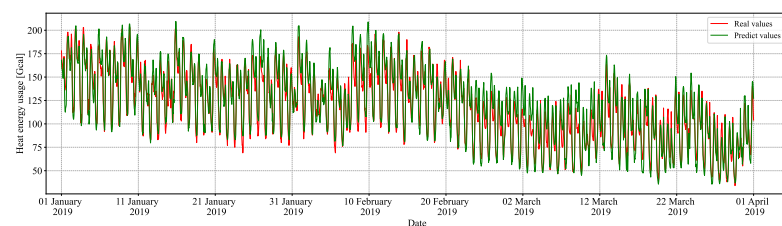
Scenario	Train Start Year	Train End Year	Test Year
A	2012	2016	2017
B	2013	2017	2018
C	2014	2018	2019
D	2015	2019	2020
E	2016	2020	2021

Table 8 displays the performance of five models across five scenarios. According to these results, the LightGBM generally performs well in five scenarios. The R2 score is high, and the MAE, RMSE, and MAPE are low. Additionally, the training time is relatively short. Therefore, LightGBM is the best option for this dataset. Catboost can also be considered as a viable alternative. The Catboost demonstrates similar performance to the LightGBM, especially in scenario C. It is noteworthy that this model performs well in terms of training time, as well as predictive performance. Although preprocessing was initially performed using the LightGBM, it is evident that this approach yields good performance for other models as well. These results indicate that the proposed optimization method is model-agnostic and can lead to an overall performance improvement.

Table 8. Performance comparison of five models in different scenarios for heat energy prediction.

Scenario	Model	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]	Time [s]
A	XGBOOST	0.9734	6.2281	8.5325	13.95	0.4
	LightGBM	0.9744	6.1804	8.3750	14.30	0.1
	Catboost	0.9747	6.0595	8.3263	13.65	2.9
	MLP	0.9662	7.0645	9.6226	16.80	251.9
	LSTM	0.9648	6.9525	9.8138	14.18	765.6
B	XGBOOST	0.9739	6.2780	8.6891	13.39	0.4
	LightGBM	0.9760	6.0781	8.3401	12.61	0.1
	Catboost	0.9755	6.0694	8.4194	12.67	2.7
	MLP	0.9662	7.2205	9.8889	16.19	213.7
	LSTM	0.9729	6.3732	8.8609	13.27	544.4
C	XGBOOST	0.9692	6.1756	8.5203	12.47	0.4
	LightGBM	0.9700	6.1441	8.4103	12.37	0.1
	Catboost	0.9728	5.8248	8.0139	11.73	2.7
	MLP	0.9545	7.5728	10.3559	15.13	208.4
	LSTM	0.9662	6.4732	8.9294	12.57	547.5
D	XGBOOST	0.9603	7.2034	9.7997	12.75	0.5
	LightGBM	0.9616	7.1646	9.6484	12.62	0.1
	Catboost	0.9651	6.7784	9.1990	12.00	2.7
	MLP	0.9644	6.9020	9.2841	12.45	205.7
	LSTM	0.9683	6.5697	8.7652	12.34	532.7
E	XGBOOST	0.9571	7.9204	11.1995	13.55	0.6
	LightGBM	0.9591	7.8274	10.9330	13.62	0.1
	Catboost	0.9592	7.6729	10.9144	12.99	2.7
	MLP	0.9579	8.4093	11.0955	17.57	252.7
	LSTM	0.9519	8.8235	11.8551	17.22	734.6

Figure 9 shows the prediction results of LightGBM on the test data. The green solid line represents the predicted values, while the red solid line denotes the real values. The x -axis and y -axis represent the dates of test data and the heat energy usage [Gcal], respectively.

**Figure 9.** Heat energy prediction results for scenario C based on LightGBM predictions using test data.

4.2. Electric Energy Dataset in Jeju, Republic of Korea

This section details the experiments conducted using the proposed architecture with both district heat energy and electric energy to evaluate its scalability across different domains. Section 4.2.1 provides an overview of the data used in these experiments, whereas Section 4.2.2 details the process optimization techniques described in Section 3.2. Finally, Section 4.2.3 delivers a comparative evaluation of five distinct models based on the outcomes from Section 4.2.2 across different scenarios.

4.2.1. Dataset Description

In addition to heat energy, a dataset from a power generation unit on Jeju Island, Republic of Korea, was used to predict actual electric energy usage. The data are collected on an hourly basis from 2007 to 2021, comprising a total of 131,496 data points. Electricity demand performance refers to the electricity demand amount adjusted to the urgently required electricity amount at the power generation unit. The unit is megawatt-hour (MWh),

which represents the amount of electric energy consumed per unit of time. Figure 10 visualizes the electric energy usage, displaying a recurring pattern with a one-year cycle.

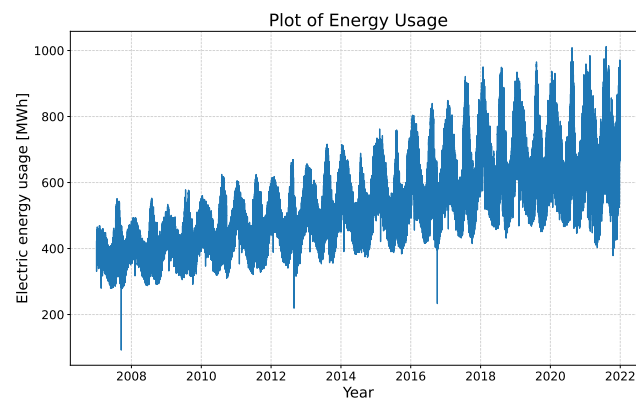


Figure 10. A plot of electric energy dataset.

To understand this repeating cycle, monthly average energy usage and temperature are visualized as a monthly plot in Figure 11. The x-axis corresponds to the months, while the left y-axis represents the monthly average electric energy usage, and the right y-axis represents the monthly average temperature in Jeju. The monthly average value is calculated by summing the values for the same month in different years and dividing it by the number of data points for that month. From this visualization, it is evident that energy usage is higher in high temperatures or low temperatures compared with others.

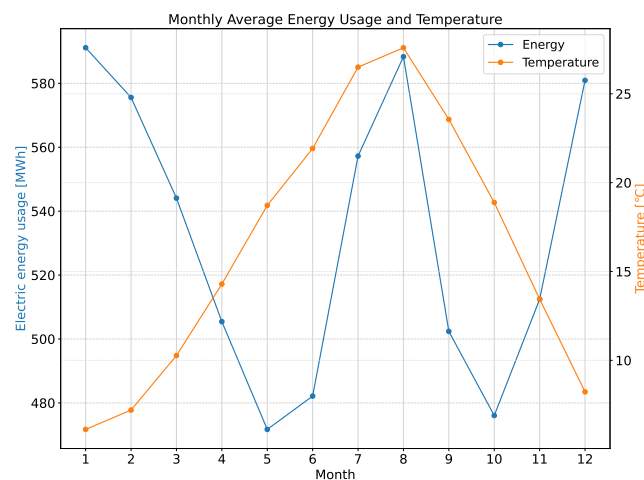


Figure 11. Monthly average electric energy usage and temperature.

Figure 12 visualizes the relationship between daily mean temperature and electric energy usage deviation. Electric energy usage deviation tends to increase at high temperatures or lower temperatures, which may be related to the intense use of electric heating and cooling devices such as air conditioners, respectively [70]. This trend indicates that there is a seasonal pattern in energy usage and shows that it is very closely related to weather.

Climate data from Jeju are used to predict electricity consumption in Jeju per hour. The latitude and longitude coordinates for Jeju are approximately 33.5141 and 126.5297. The minimum, maximum, average, and standard deviation of the meteorological variables and electric energy used for prediction can be found in Table 9.

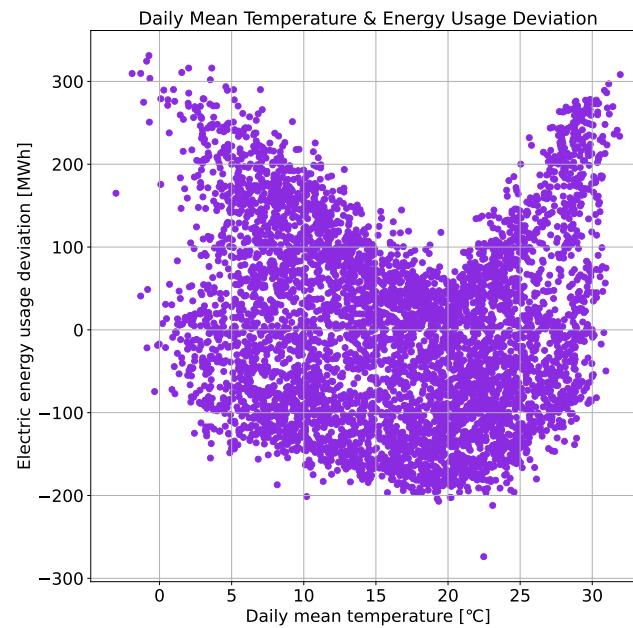


Figure 12. Daily mean temperature and electric energy usage deviation.

Table 9. The minimum, maximum, mean, and standard deviation of meteorological variables and electric energy for Jeju.

Name	Minimum	Maximum	Mean	Standard	Unit
Datetime	1 January 2007	31 December 2021	-	-	-
Temperature	−5.3	36.8	16.44	7.96	°C
Wind speed	0	26.6	3.23	1.91	m/s
Wind direction	0	360	197.54	116.42	16 directions
Humidity	7	100	68.47	15.52	%
Dew-point temperature	−18.8	28.3	10.49	9.29	°C
Local atmospheric pressure	972.7	1036.4	1013.92	7.9	hPa
Sunshine duration	0	1	0.38	0.43	h
Solar radiation	0	4.09	1.03	1	MJ/m ²
Visibility	6	5553	1754.03	650.58	10 m
Ground temperature	−3.8	66.8	18.16	11.1	°C
Energy	92.30	1012.10	532.31	124.91	MWh

4.2.2. Process Optimization Methods

In the process optimization methods, a total of 3360 scenarios are generated based on conditions. These scenarios are generated by combining 8 different datasets, 4 data split patterns, and 105 different data splits of ratio and year. Table 10 shows the data split ratio of 3360 scenarios.

Data split patterns were grouped in the same way as the process optimization method of heat energy, and 10% of the high-performance data were sorted and combined within each group and metric. As a result of comprehensively considering the distribution of data cleaning methods in the top 10% of sorted data, it was determined that dataset3 was to be examined.

Then, the data cleaning was grouped, and the distribution of high-performance data split patterns was examined to select the data split patterns. As a result, when the data split pattern was 12 months, the overall performance was high.

Table 10. The train-to-test ratio of training data and test data for electric energy prediction.

Train Ratio	Test Ratio	Count	Percentage of Count [%]
50.0	50.0	448	13.3
66.7	33.3	416	12.4
75.0	25.0	384	11.4
80.0	20.0	352	10.5
83.3	16.7	320	9.5
85.7	14.3	288	8.6
87.5	12.5	256	7.6
88.9	11.1	224	6.7
90.0	10.0	192	5.7
90.9	9.1	160	4.8
91.7	8.3	128	3.8
92.3	7.7	96	2.9
92.9	7.1	64	1.9
93.3	6.7	32	1.0
Total		3360	100

Finally, the data split ratio was determined by grouping based on the previously selected dataset3 and 12 months. Based on the sorted top data, the data split ratio that demonstrates the highest performance is identified. Despite the differential training of data split ratios, it can be confirmed that the case trained with a ratio of 87.5:12.5 exhibits the highest performance.

Table 11 presents the maximum, minimum, and average values for each indicator in 3360 cases, obtained by combining eight data cleaning methods, four data split pattern methods, and 105 methods that consider the data split ratio and frequency for electric energy. For the entire case, the difference between the maximum and minimum values for each average indicator is 0.7494 for R2 score, 32.0729 for MAE, 40.0391 for RMSE, and 4.29 for MAPE. For 0.8868 with a maximum value based on the R2 score in the total number of cases, MAE, RMSE, and MAPE did not demonstrate the best performance, at 24.0195, 32.8413, and 3.81, respectively, and when MAPE had the best performance at 3.80, R2 score, MAE, and RMSE did not have the best performance, at 0.8827, 24.0023, and 33.4396, respectively. If the MAE had the highest performance at 19.9757, the RMSE also had the highest performance at 25.1399, but the R2 score and MAPE at this time were 0.7142 and 4.83, respectively, which were not the highest performance values. This result was the same as heat energy, and it was confirmed once again that it is important to consider evenly because the optimization method of the model may vary depending on which indicator is focused. Table 12 shows the maximum, minimum, and average values of the performance of each evaluation index for cases that satisfy the data cleaning method and the data split pattern method selected through a proposed method that evenly considers the used evaluation indicators. While the highest performance values for each indicator in Table 12 may differ from those in Table 11, a comparison of the average values shows overall improved performance.

Table 11. Categorized cases performance summary for electric energy dataset.

Metrics	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]
Max	0.8868	52.0486	65.1790	8.09
Mean	0.7279	34.4093	42.9820	5.74
Min	0.1374	19.9757	25.1399	3.80
Count	3360	3360	3360	3360

Furthermore, a comparison of the results obtained through the proposed method with those obtained from training all but the same test data revealed that the R2 score was improved by 0.0098, the MAE was 1.1626, and the RMSE was improved by 0.17. This result

demonstrates that the use of an appropriate amount of data for the analysis of time series patterns, such as thermal energy, can result in satisfactory performance.

Table 12. A set of cases performance summary for electric energy dataset which satisfied selected data cleaning and data split pattern.

Metrics	R2 Score	MAE [Gcal]	RMSE [Gcal]	MAPE [%]
Max	0.8868	42.6838	52.6884	6.74
Mean	0.7743	32.3221	40.0210	5.41
Min	0.6114	20.7920	25.7141	3.81
Count	105	105	105	105

4.2.3. Prediction

In nine different scenarios, five models compared their performance by using optimized data. As indicated in Table 13, scenarios were trained in a different year, and the other condition was determined by the process optimization method.

Table 13. The period of each scenario for electric energy prediction.

Scenario	Train Start Year	Train End Year	Test Year
A	2007	2012	2013
B	2008	2013	2014
C	2009	2014	2015
D	2010	2015	2016
E	2011	2016	2017
F	2012	2017	2018
G	2013	2018	2019
H	2014	2019	2020
I	2015	2020	2021

Table 14 shows the performance of five models in nine scenarios. Each model excels in various aspects, making the selection of the final model a challenging task. The LSTM demonstrates remarkable R2 score performance in some scenarios, particularly in modeling sequential data like time series. However, it incurs relatively long training times. MLP also excels in multiple scenarios in terms of MAE and RMSE, showcasing its ability to model nonlinear relationships. Nonetheless, they can be time-consuming. Additionally, boosting models such as Catboost, LightGBM, and XGBOOST are capable of rapid learning, with Catboost excelling in R2 score, MAE, and RMSE in certain scenarios. Taking the experimental results into comprehensive consideration, Catboost emerges as the preferred choice as the final model for the given dataset and scenarios. These results indicate that the proposed architecture improves the performance not only in heat energy but also in electric energy, opening up possibilities for extending the architecture to other energy predictions.

Figure 13 illustrates the electric energy prediction results for scenario I using LightGBM on test data. The green solid line denotes the predicted values, whereas the red solid line represents the observed values. The x -axis displays the dates of the test data, and the y -axis represents the electric energy usage [MWh].

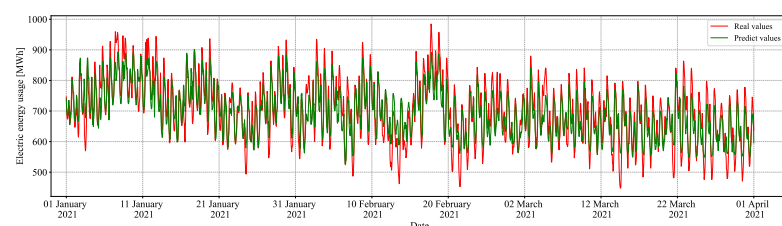


Figure 13. Electric energy prediction results for scenario I based on LightGBM predictions using test data.

Table 14. Performance comparison of five models in different scenarios for electric energy prediction.

Scenario	Model	R2 Score	MAE [MWh]	RMSE [MWh]	MAPE [%]	Time [s]
A	XGBOOST	0.7309	30.8266	36.0503	5.95	0.4
	LightGBM	0.7154	31.7121	37.0707	6.08	0.1
	Catboost	0.7396	30.7598	35.4616	5.92	2.8
	MLP	0.8521	20.1635	26.7258	3.94	255.0
	LSTM	0.8313	21.9466	28.5437	4.22	659.2
B	XGBOOST	0.6730	32.6795	38.5980	6.05	0.4
	LightGBM	0.6560	33.3752	39.5839	6.14	0.2
	Catboost	0.6959	31.5410	37.2187	5.84	2.9
	MLP	0.8437	20.1325	26.6869	3.84	250.0
	LSTM	0.8714	18.3526	24.2021	3.48	649.6
C	XGBOOST	0.7543	30.4345	37.1846	5.37	0.4
	LightGBM	0.7679	29.7199	36.1433	5.23	0.1
	Catboost	0.7726	29.4492	35.7776	5.19	3.2
	MLP	0.8691	20.1768	27.1443	3.65	305.5
	LSTM	0.8207	24.3074	31.7660	4.51	929.2
D	XGBOOST	0.7410	35.5428	44.7229	5.86	0.4
	LightGBM	0.7487	35.4434	44.0572	5.82	0.1
	Catboost	0.7594	34.8588	43.1042	5.75	3.0
	MLP	0.8381	26.0602	35.3564	4.33	301.0
	LSTM	0.8602	24.1559	32.8527	4.08	894.6
E	XGBOOST	0.7727	38.1528	45.7795	5.99	0.4
	LightGBM	0.7879	36.6678	44.2164	5.73	0.1
	Catboost	0.7826	37.5471	44.7658	5.89	3.1
	MLP	0.8456	28.5964	37.7340	4.68	265.5
	LSTM	0.8907	23.8404	31.7486	3.82	652.4
F	XGBOOST	0.8109	35.9928	44.4130	5.43	0.4
	LightGBM	0.8109	36.1888	44.4181	5.43	0.1
	Catboost	0.8249	34.6577	42.7370	5.22	3.0
	MLP	0.8735	27.2717	36.3355	4.30	255.4
	LSTM	0.8880	25.3480	34.1788	3.93	651.5
G	XGBOOST	0.8549	29.0200	36.7715	4.37	0.4
	LightGBM	0.8491	29.4623	37.4979	4.40	0.1
	Catboost	0.8619	28.2419	35.8748	4.25	3.0
	MLP	0.8054	32.4475	42.5840	5.18	301.7
	LSTM	0.8908	23.2536	31.8907	3.60	910.5
H	XGBOOST	0.8766	25.1140	34.3171	3.96	0.4
	LightGBM	0.8824	24.4052	33.5056	3.87	0.1
	Catboost	0.8885	23.5458	32.6237	3.74	2.9
	MLP	0.7968	32.5388	44.0326	5.18	301.8
	LSTM	0.8239	30.9385	40.9914	4.95	893.0
I	XGBOOST	0.7895	39.2159	48.2842	6.08	0.4
	LightGBM	0.7887	39.5257	48.3825	6.11	0.1
	Catboost	0.7986	38.8042	47.2304	6.01	2.9
	MLP	0.5811	58.2610	68.1151	8.72	244.4
	LSTM	0.7272	47.1241	54.9711	7.19	652.2

4.3. Prediction Results and Discussions

Empirical results using heat energy and electric energy confirmed the following: data cleaning, data split patterns, and data split ratios, selected as conditions for optimizing data, significantly impacted model performance. This underscores the importance of selecting conditions tailored to the specific data. Additionally, high performance across various indicators was achieved by considering both the characteristics of the indicators and the inherent properties of time series data.

This section describes the input variables that influenced the model's results through the application of XAI. The prediction models of heat energy and electric energy are explained, respectively, and compared and analyzed for the two domains. Figures 14 and 15 visualize the importance of variables using SHAP for a LightGBM applied to predict heat energy and electric energy. In a summary plot generated using SHAP values, the *y*-axis displays the importance of features in descending order, with the most important features at the top and the least important at the bottom. And the *x*-axis shows the SHAP values, indicating their contributions to increasing the output when positive and decreasing the output when negative.

In heat energy, the top five important variables selected are ground temperature, temperature, hour, month, and solar radiation (Radiation). Figure 14 explains that high values of ground temperature and temperature contribute to predicting lower energy usage, and lower values contribute to predicting higher energy usage. In the case of time, it contributes to predicting lower energy usage when it has low values and higher energy usage when it has a high value.

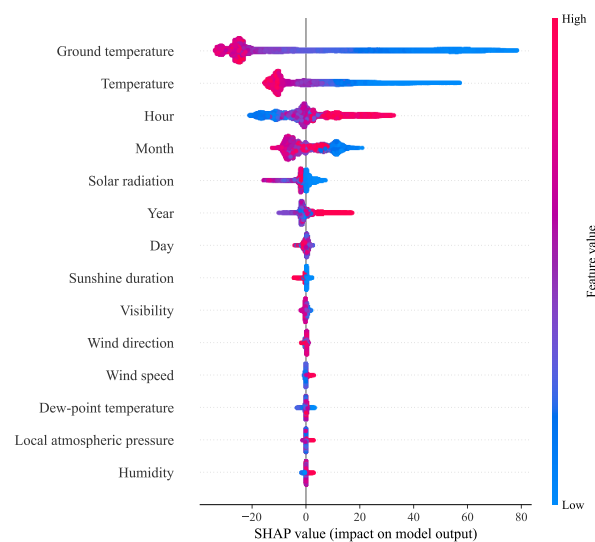


Figure 14. SHAP explanation for heat energy prediction.

On the other hand, in electric energy, the top five important variables selected are year, hour, temperature, ground temperature, and month. Figure 15 explains that the low values of Year and Hour contribute to predicting lower energy usage, and high values contribute to predicting higher energy usage. In the case of ground temperature and temperature, it is explained that if the value is too low or too high, it contributes to predicting more energy usage.

These indicate that different features have a significant impact on the model depending on the dataset. It can be interpreted that energy types like heating are more influenced by temperature, while electric energy is interpreted to be greatly influenced by time. These results suggest that seasonal and social issues affect energy prediction such as temperature, time, and day.

The primary factors influencing energy output were identified through the XAI methodology. The analysis revealed that time and temperature variables significantly impact energy output. SHAP values provided detailed insights into the model's decision-making process by quantifying each feature's contribution. Elucidating these influential factors through XAI clarifies how the model makes predictions and why. This enhances the model's robustness and predictive accuracy. Furthermore, this approach can be applied to other tasks, such as dimensionality reduction or feature selection, thus improving overall model performance and reliability in various applications.

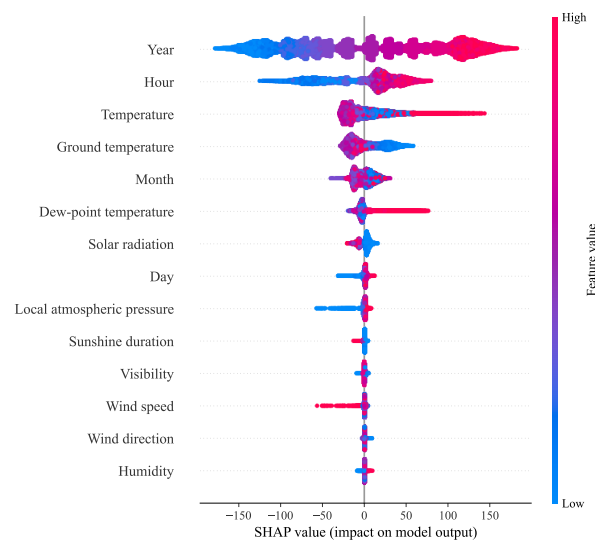


Figure 15. SHAP explanation for electric energy prediction.

Two experiments have confirmed that the optimization data determination method in the preprocessing phase leads to performance improvement. Choosing a low-carbon method for energy production by accurately predicting energy consumption in advance with improved performance can be an effective way to achieve carbon neutrality. In addition, despite the different patterns of heat and electric energy, the proposed architecture has been confirmed to be effective not only in heat energy but also in electric energy. These results can be expected to have scalability that can be effectively applied to other energy data affected by climate, lifestyle, and time. If energy consumption is predicted by effectively utilizing various energy data, energy management and optimization will enable energy conservation and sustainable energy use directions to be planned. Besides, it is expected to help with energy management and optimization by providing a way to build and interpret a data-based energy prediction model.

In this study, the preprocessing of time series energy data and the use of SHAP were employed to develop a robust, generalizable model. Key features with significant temporal changes and high fluctuation likelihood were carefully considered to ensure model reliability. Additionally, the model is designed to utilize four metrics to assess its performance. While this provides a robust foundation for evaluation, future research could involve the incorporation of additional evaluation metrics. This would enhance the assessment process, providing a more comprehensive understanding of the method's performance across different conditions and datasets.

The approach proposed by the authors emphasizes enhancing the model's generalizability through comprehensive experimental design and robust preprocessing techniques. Methods such as MinMaxScaler for normalization, advanced imputation strategies for handling missing data, and careful feature engineering were incorporated to ensure that the model could effectively adapt to various time series datasets. These preprocessing steps help maintain the integrity and consistency of the data, which is crucial for achieving reliable predictions across different contexts. Furthermore, the model was evaluated using a diverse set of prediction algorithms, including XGBoost, LightGBM, CatBoost, MLP, and LSTM. This multimodel evaluation demonstrated that the proposed approach consistently improves performance, regardless of the specific algorithm used. The use of SHAP values for explainability also contributed to the model's robustness by identifying and emphasizing the most influential features, thereby enhancing the interpretability and reliability of the predictions. Experimental results indicate that the proposed method not only achieves high predictive accuracy but also maintains strong generalizability across different datasets and conditions. This is evidenced by the consistent performance improvements observed in various test scenarios, highlighting the adaptability of the proposed approach. However,

certain limitations, such as sensitivity to extreme outliers and domain-specific nuances, are acknowledged and may require further attention in future research.

A superior model performs well under all circumstances. It is crucial to investigate models that demonstrate exceptional performance and adaptability to sudden data changes. Future developments will focus on generating virtual data representing anomalous phenomena using generative adversarial networks or other generative models. This approach aims to create models capable of simulating and adapting to unusual events, thereby enhancing their robustness and applicability.

5. Conclusions

This paper focuses on exploring effective preprocessing techniques and proposes a time series prediction architecture based on these techniques. In time series prediction, to enhance the accuracy of predicting results, three aspects are considered when finding the optimal data optimization methods: data cleaning, data split patterns, and data split ratios. Data cleaning considers potential issues with missing data that may arise in time series data analysis. Data split patterns consider characteristics such as temporal dependencies, seasonality, and periodicity, inherent in time series data.

Lastly, data split ratios are determined by considering factors such as temporal issues, trends, and the amount of data. Empirical results show performance improvements achieved through data optimization and confirm the importance of data preprocessing methods. The evaluation of five different prediction models confirmed that the proposed preprocessing method enhances predictive results, regardless of the model employed. In the final step, XAI was applied to the prediction results to analyze the factors influencing energy prediction. Furthermore, electric energy prediction experiments were performed to confirm the improvement in prediction performance and to show the scalability to other types of energy prediction by using the same input variables as heat energy.

An enhanced energy prediction method enables the determination of required production volumes in advance and the selection of appropriate low-carbon energy production methods for supply. Future research will focus on extending the proposed architecture to more advanced models, such as hybrid deep learning frameworks, and exploring its application in different geographical regions and energy systems. In addition, incorporating real-time data from Internet of Things (IoT) sensors and investigating the integration of renewable energy sources in energy forecasting will be key areas for further study.

Author Contributions: Conceptualization, J.S., H.M. and S.L.; investigation, C.-J.C. and T.S.; writing—original draft, J.S. and S.L.; writing—review and editing, J.S., H.M. and S.L.; visualization, E.K.; supervision, H.M. and S.L.; project administration, H.M. and S.L.; funding acquisition, H.M. and E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2023-RS-2023-00254529) grant funded by the Korean government (MSIT), and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A6A3A01078538), and by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00106, Development of explainable AI-based diagnosis and analysis framework using energy demand big data in multiple domains).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to express their sincere gratitude to Hyoung-Kyu Song, for his invaluable support and guidance throughout the course of this research. His contributions, particularly in providing the necessary funding and insightful advice, were instrumental to the completion of this project.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DHS	District heating system
ML	Machine learning
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting
LightGBM	Light gradient-boosting machine
CatBoost	Categorical boosting
MLP	Multilayer perceptron
LSTM	Long short-term memory
GBDT	Gradient boosting decision trees
SVM	Support vector machines
SHAP	Shapley additive explanations
R2 score	R-squared or coefficient of determination
MAE	Mean absolute error
MSE	Mean square error
RMSE	Root mean square error
MAPE	Mean absolute percentage error
Gcal	Gigacalories
MWh	Megawatt-hour
IoT	Internet of Things

References

1. Lee, H.; Calvin, K.; Dasgupta, D.; Krinner, G.; Mukherji, A.; Thorne, P.; Trisos, C.; Romero, J.; Aldunce, P.; Barrett, K.; et al. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneva, Switzerland, 2023.
2. Masson-Delmotte, V.; Zhai, P.; Pörtner, H.; Roberts, D.; Skea, J.; Shukla, P.; Pirani, A.; Moufouma-Okia, W.; Péan, C.; Pidcoc, R.; et al. Intergovernmental Panel on Climate Change (IPCC). Global Warming of 1.5 °C: An IPCC Special Report on the Impacts of Global Warming of 1.5 °C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change. Sustainable Development, and Efforts to Eradicate Poverty. 2018. Available online: <https://www.ipcc.ch/sr15/> (accessed on 18 October 2019).
3. Rezaie, B.; Rosen, M.A. District heating and cooling: Review of technology and potential enhancements. *Appl. Energy* **2012**, *93*, 2–10. [CrossRef]
4. Vakhnin, A.; Ryzhikov, I.; Brester, C.; Niska, H.; Kolehmainen, M. Weather-Based Prediction of Power Consumption in District Heating Network: Case Study in Finland. *Energies* **2024**, *17*, 2840. [CrossRef]
5. Yang, Y.; Fan, C.; Xiong, H. A novel general-purpose hybrid model for time series forecasting. *Appl. Intell.* **2022**, *52*, 2212–2223.
6. Liu, Z.; Zhu, Z.; Gao, J.; Xu, C. Forecast Methods for Time Series Data: A Survey. *IEEE Access* **2021**, *9*, 91896–91912. [CrossRef]
7. Xue, P.; Jiang, Y.; Zhou, Z.; Chen, X.; Fang, X.; Liu, J. Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms. *Energy* **2019**, *188*, 116085.
8. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; KDD'16; pp. 785–794. [CrossRef]
9. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
10. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
11. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
12. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
13. Xiao, Z.; Gang, W.; Yuan, J.; Chen, Z.; Li, J.; Wang, X.; Feng, X. Impacts of data preprocessing and selection on energy consumption prediction model of HVAC systems based on deep learning. *Energy Build.* **2022**, *258*, 111832.
14. Zhou, Y.; Liu, Y.; Wang, D.; Liu, X. Comparison of machine-learning models for predicting short-term building heating load using operational parameters. *Energy Build.* **2021**, *253*, 111505. [CrossRef]
15. Runge, J.; Saloux, E. A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system. *Energy* **2023**, *269*, 126661.
16. Dang, L.M.; Lee, S.; Li, Y.; Oh, C.; Nguyen, T.N.; Song, H.K.; Moon, H. Daily and seasonal heat usage patterns analysis in heat networks. *Sci. Rep.* **2022**, *12*, 9165. [CrossRef] [PubMed]

17. Başağaoğlu, H.; Chakraborty, D.; Lago, C.D.; Gutierrez, L.; Şahinli, M.A.; Giacomoni, M.; Furl, C.; Mirchi, A.; Moriasi, D.; Şengör, S.S. A review on interpretable and explainable artificial intelligence in hydroclimatic applications. *Water* **2022**, *14*, 1230. [\[CrossRef\]](#)
18. Lin, J.; Lin, W.; Lin, W.; Wang, J.; Jiang, H. Thermal prediction for air-cooled data center using data driven-based model. *Appl. Therm. Eng.* **2022**, *217*, 119207. [\[CrossRef\]](#)
19. Lin, T.; Pan, Y.; Xue, G.; Song, J.; Qi, C. A novel hybrid spatial-temporal attention-LSTM model for heat load prediction. *IEEE Access* **2020**, *8*, 159182–159195. [\[CrossRef\]](#)
20. Leiprecht, S.; Behrens, F.; Faber, T.; Finkenrath, M. A comprehensive thermal load forecasting analysis based on machine learning algorithms. *Energy Rep.* **2021**, *7*, 319–326. [\[CrossRef\]](#)
21. Hou, C.; Wu, J.; Cao, B.; Fan, J. A deep-learning prediction model for imbalanced time series data forecasting. *Big Data Min. Anal.* **2021**, *4*, 266–278. [\[CrossRef\]](#)
22. Noussan, M.; Jarre, M.; Poggio, A. Real operation data analysis on district heating load patterns. *Energy* **2017**, *129*, 70–78. [\[CrossRef\]](#)
23. Kim, S.; Song, Y.; Sung, Y.; Seo, D. Development of a consecutive occupancy estimation framework for improving the energy demand prediction performance of building energy modeling tools. *Energies* **2019**, *12*, 433. [\[CrossRef\]](#)
24. Chakraborty, D.; Alam, A.; Chaudhuri, S.; Başağaoğlu, H.; Sulbaran, T.; Langar, S. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. *Appl. Energy* **2021**, *291*, 116807. [\[CrossRef\]](#)
25. Chung, W.J.; Liu, C. Analysis of input parameters for deep learning-based load prediction for office buildings in different climate zones using eXplainable Artificial Intelligence. *Energy Build.* **2022**, *276*, 112521. [\[CrossRef\]](#)
26. Sim, T.; Choi, S.; Kim, Y.; Youn, S.H.; Jang, D.J.; Lee, S.; Chun, C.J. eXplainable AI (XAI)-Based Input Variable Selection Methodology for Forecasting Energy Consumption. *Electronics* **2022**, *11*, 2947. [\[CrossRef\]](#)
27. Chou, J.S.; Truong, D.N. Multistep energy consumption forecasting by metaheuristic optimization of time-series analysis and machine learning. *Int. J. Energy Res.* **2021**, *45*, 4581–4612. [\[CrossRef\]](#)
28. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
29. Guelpa, E.; Marincioni, L.; Capone, M.; Deputato, S.; Verda, V. Thermal load prediction in district heating systems. *Energy* **2019**, *176*, 693–703. [\[CrossRef\]](#)
30. Brownlee, J. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*; Machine Learning Mastery: San Juan, Puerto Rico, 2018.
31. Barrera-Animas, A.Y.; Oyedele, L.O.; Bilal, M.; Akinosho, T.D.; Delgado, J.M.D.; Akanbi, L.A. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Mach. Learn. Appl.* **2022**, *7*, 100204. [\[CrossRef\]](#)
32. Kim, H.U.; Bae, T.S. Preliminary study of deep learning-based precipitation prediction. *J. Korean Soc. Surv. Geod. Photogramm. Cartogr.* **2017**, *35*, 423–429.
33. Shanker, M.; Hu, M.Y.; Hung, M.S. Effect of data standardization on neural network training. *Omega* **1996**, *24*, 385–397. [\[CrossRef\]](#)
34. Alkhayat, G.; Mehmood, R. A review and taxonomy of wind and solar energy forecasting methods based on deep learning. *Energy AI* **2021**, *4*, 100060. [\[CrossRef\]](#)
35. Hoffmann, M.; Kotzur, L.; Stolten, D.; Robinius, M. A review on time series aggregation methods for energy system models. *Energies* **2020**, *13*, 641. [\[CrossRef\]](#)
36. Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M.A. Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting. *Energies* **2020**, *13*, 391. [\[CrossRef\]](#)
37. Lara-Benítez, P.; Carranza-García, M.; Luna-Romera, J.M.; Riquelme, J.C. Temporal convolutional networks applied to energy-related time series forecasting. *Appl. Sci.* **2020**, *10*, 2322. [\[CrossRef\]](#)
38. Khan, W.; Walker, S.; Zeiler, W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy* **2022**, *240*, 122812. [\[CrossRef\]](#)
39. Mallapragada, D.S.; Papageorgiou, D.J.; Venkatesh, A.; Lara, C.L.; Grossmann, I.E. Impact of model resolution on scenario outcomes for electricity sector system expansion. *Energy* **2018**, *163*, 1231–1244. [\[CrossRef\]](#)
40. Huang, C.J.; Shen, Y.; Chen, Y.H.; Chen, H.C. A novel hybrid deep neural network model for short-term electricity price forecasting. *Int. J. Energy Res.* **2021**, *45*, 2511–2532. [\[CrossRef\]](#)
41. de Amorim, L.B.; Cavalcanti, G.D.; Cruz, R.M. The choice of scaling technique matters for classification performance. *Appl. Soft Comput.* **2023**, *133*, 109924. [\[CrossRef\]](#)
42. Raju, V.N.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; pp. 729–735. [\[CrossRef\]](#)
43. LeCun, Y.; Bottou, L.; Orr, G.B.; Müller, K.R. Efficient backprop. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 9–50.
44. Huang, L.; Qin, J.; Zhou, Y.; Zhu, F.; Liu, L.; Shao, L. Normalization Techniques in Training DNNs: Methodology, Analysis and Application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10173–10196. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

46. Yozgatligil, C.; Aslan, S.; Iyigun, C.; Batmaz, I. Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. *Theor. Appl. Climatol.* **2013**, *112*, 143–167. [\[CrossRef\]](#)
47. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793.
48. Fan, C.; Chen, M.; Wang, X.; Wang, J.; Huang, B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front. Energy Res.* **2021**, *9*, 652801. [\[CrossRef\]](#)
49. Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; Tabona, O. A survey on missing data in machine learning. *J. Big Data* **2021**, *8*, 140. [\[CrossRef\]](#)
50. Khan, S.I.; Hoque, A.S.M.L. SICE: An improved missing data imputation technique. *J. Big Data* **2020**, *7*, 37. [\[CrossRef\]](#)
51. Weerakody, P.B.; Wong, K.W.; Wang, G.; Ela, W. A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing* **2021**, *441*, 161–178. [\[CrossRef\]](#)
52. Chen, M.; Zhu, H.; Chen, Y.; Wang, Y. A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere* **2022**, *13*, 1044. [\[CrossRef\]](#)
53. Mudassir, M.; Bennbaia, S.; Unal, D.; Hammoudeh, M. Time-series forecasting of Bitcoin prices using high-dimensional features: A machine learning approach. *Neural Comput. Appl.* **2020**, 1–15. [\[CrossRef\]](#)
54. Nguyen, X.H. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. *Adv. Water Resour.* **2020**, *142*, 103656.
55. Menezes, A.G.; Mastelini, S.M. MegazordNet: Combining statistical and machine learning standpoints for time series forecasting. *arXiv* **2021**, arXiv:2107.01017.
56. Ye, J.; Zhao, B.; Deng, H. Photovoltaic Power Prediction Model Using Pre-train and Fine-tune Paradigm Based on LightGBM and XGBoost. *Procedia Comput. Sci.* **2023**, *224*, 407–412. [\[CrossRef\]](#)
57. Aksoy, N.; Genc, I. Predictive models development using gradient boosting based methods for solar power plants. *J. Comput. Sci.* **2023**, *67*, 101958. [\[CrossRef\]](#)
58. Zhu, X.; Shen, X.; Chen, K.; Zhang, Z. Research on the prediction and influencing factors of heavy duty truck fuel consumption based on LightGBM. *Energy* **2024**, *296*, 131221. [\[CrossRef\]](#)
59. Chola, A.; Rastogi, R.; Kaur, P.; Chaudhary, A.; Biswas, D. Predictive Analytics Beyond the Hype: A Comprehensive Comparison of LSTM, XGBoost and LightGBM with Emphasis on RMSE and CPU Utilization. In Proceedings of the 2024 Third International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 18–20 January 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 179–184.
60. Haque, H.; Razzak, M.A. Medium-term Energy Demand Analysis using Machine Learning: A Case Study on a Sub-District Area of a Divisional City in Bangladesh. *IEEE Trans. Ind. Appl.* **2024**, *60*, 4424–4432. [\[CrossRef\]](#)
61. Sasikala, D.; Theetchenya, S. A Comparative Exploration of Time Series Models for Wild Fire Prediction. In Proceedings of the 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 11–12 January 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5.
62. Antypas, E.; Spanos, G.; Lalas, A.; Votis, K.; Tzovaras, D. A time-series approach for estimated time of arrival prediction in autonomous vehicles. *Transp. Res. Procedia* **2024**, *78*, 166–173. [\[CrossRef\]](#)
63. Hu, B.; Palta, M.; Shao, J. Properties of R2 statistics for logistic regression. *Stat. Med.* **2006**, *25*, 1383–1395. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [\[CrossRef\]](#)
66. Hussain, L.; Saeed, S.; Idris, A.; Awan, I.A.; Shah, S.A.; Majid, A.; Ahmed, B.; Chaudhary, Q.A. Regression analysis for detecting epileptic seizure with different feature extracting strategies. *Biomed. Eng./Biomed. Tech.* **2019**, *64*, 619–642. [\[CrossRef\]](#)
67. de Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* **2016**, *192*, 38–48. [\[CrossRef\]](#)
68. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [\[CrossRef\]](#)
69. Soutullo, S.; Bujedo, L.A.; Samaniego, J.; Borge, D.; Ferrer, J.A.; Carazo, R.F.; del Rosario Heras, M. Energy performance assessment of a polygeneration plant in different weather conditions through simulation tools. *Energy Build.* **2016**, *124*, 7–18. [\[CrossRef\]](#)
70. Apadula, F.; Bassini, A.; Elli, A.; Scapin, S. Relationships between meteorological variables and monthly electricity demand. *Appl. Energy* **2012**, *98*, 346–356. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.