IJRCS

International Journal of Robotics and Control Systems
Vol. 4, No. 1, 2024, pp. 33-49
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

ASCEE

Predictive Modeling of Energy Consumption in the Steel Industry Using CatBoost Regression: A Data-Driven Approach for Sustainable Energy Management

K. Karthick [a,1,*], R. Dharmaprakash [b,2], S. Sathya [c,3]

[a] Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam - 532127, Andhra Pradesh, India
[b] Department of Electrical and Electronics Engineering, Panimalar Engineering College, Chennai – 600123, Tamil Nadu, India
[c] Department of Electrical and Electronics Engineering, S. A. Engineering College, Chennai-600077, Tamil Nadu, India
[1] kkarthiks@gmail.com; [2] rdharmaprakash@yahoo.co.in; [3] sathya_vrscet@yahoo.co.in
* Corresponding Author

ARTICLE INFO

Article history
Received October 11, 2023
Revised November 24, 2023
Accepted January 03, 2024

Keywords
Energy Consumption;
Data Analysis;
Predictive Modeling;
Machine Learning in Steel Industry;
Energy Optimization

ABSTRACT

This article presents a machine learning model for predicting energy consumption in the steel industry, which aids in energy management, cost reduction, environmental regulation compliance, informed decision-making for future energy investments, and contributes to sustainability. The dataset used for the prediction model comprises 11 attributes and 35,040 instances. The CatBoost prediction algorithm was employed for energy consumption prediction, and hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation. The developed model has undergone a comparative analysis based on both Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics, demonstrating its promise for accurate energy consumption prediction on both the training and test sets. The proposed model accurately predicts energy consumption for different load types, achieving impressive results on both the training set (RMSE=0.382, R^2=0.999, MAPE=1.139) and the test set (RMSE=1.073, R^2=0.998, MAPE=1.142). These findings highlight the potential of CatBoost as a valuable tool for energy management and conservation, enabling organizations to make informed decisions, optimize resource allocation, and promote sustainability.

1. Introduction

Predicting energy usage in the steel manufacturing industry is essential for improving energy efficiency, reducing costs, meeting environmental regulations, and promoting sustainability. Energy consumption in the steel industry refers to the amount of electricity used by steel plants and facilities to produce steel products. This energy is consumed in various forms, including electricity, fuel, and heat [1]. Predicting energy usage in the steel manufacturing industry is important for several reasons. Firstly, it can help steel companies to manage their energy use and reduce costs. By accurately predicting energy consumption, steel plants can optimize their production processes and reduce energy waste, which can result in significant cost savings [2]. Secondly, predicting energy consumption can

https://doi.org/10.31763/ijrcs.v4i1.1234 ijrcs@ascee.org

help steel companies to meet environmental regulations and minimize their environmental impact by lowering carbon emissions. The steel industry is a significant source of global warming gas emissions, and minimizing energy consumption is a key part of efforts to mitigate climate change [3]. Thirdly, predicting energy consumption can help steel companies to make informed decisions about future energy investments. By understanding their energy consumption patterns and requirements, steel companies can identify opportunities to invest in renewable energy sources or energy-efficient technologies, which can further reduce costs and improve their environmental impact [4]. Machine learning falls under the umbrella of artificial intelligence, which empowers computers to learn and enhance through experience, without requiring explicit programming [5]. In the context of predicting energy usage in the steel manufacturing industry, machine learning algorithms can analyze historical data on energy usage patterns and identify complex relationships between variables that may not be obvious to humans [6]. This can help to develop accurate predictive models suitable for predicting future energy consumption and optimize energy usage in real-time. By using machine learning, it is possible to identify opportunities for energy savings, reduce energy waste, and ultimately decrease energy costs, while at the same time contributing to a more sustainable future.

Compared to conventional methods, machine learning excels in predicting energy consumption due to its ability to handle complex and nonlinear relationships between variables, and its adaptability to changing conditions and data patterns.

Conventional methods for predicting energy consumption are Regression analysis, Time-series analysis and Expert systems. Statistical techniques are employed in regression analysis [7] to model the correlation between independent and dependent variables. However, it assumes a linear relationship between variables and may not capture complex patterns.

Utilizing time-series analysis [8], this method relies on historical data to predict future energy consumption. Nevertheless, its effectiveness may be limited in adapting to changing conditions and capturing nonlinear relationships. Expert systems method relies on expert knowledge and rules to predict energy consumption. However, it may be limited by the availability of expert knowledge and may not capture all relevant variables [9]. Machine learning can automatically learn from data and adapt to changing conditions. It can handle complex and nonlinear relationships between variables, and can use multiple sources of data to make accurate predictions [10].

The proposed energy consumption prediction model for predicting the energy consumption in the steel industry can have several benefits. First, it can help the company better manage its energy use and reduce costs by identifying areas where energy efficiency can be improved. Second, it can help the company meet environmental regulations and reduce its carbon footprint by identifying areas where energy use can be reduced. Third, it can help the company make more informed decisions about future energy investments by providing a clearer picture of its current energy use and future needs. Finally, by reducing energy consumption, the company can also contribute to a more sustainable future and help address the global challenge of climate change.

The major contributions of this research are as follows:

- The research presents results that demonstrate accurate energy consumption prediction using the CatBoost regression algorithm.
- This accurate prediction of energy consumption leads to environmental benefits by reducing carbon emissions and promoting sustainable practices in the steel industry.
- The article provides valuable insights of the data that leads to significant cost savings and improved energy efficiency, aligning with the financial and environmental goals of the company.
- The use of CatBoost as the regression algorithm is a notable contribution. CatBoost efficiently handles both numerical and categorical features, making it well-suited for the energy consumption prediction task, where diverse types of data need to be analyzed for accurate predictions.

- The research applies GridSearchCV with 5-fold cross-validation for hyperparameter optimization to the CatBoost model. This optimization process fine-tunes crucial hyperparameters, such as the number of trees (iterations), tree depth, and learning rate, to achieve the best model performance. By tuning these hyperparameters, the model's predictive capabilities are improved, leading to better energy consumption predictions and overall model effectiveness.
- The proposed CatBoost approach is compared with existing methods and the results demonstrate that CatBoost outperforms these traditional algorithms in terms of various evaluation metrics. The feature importance plot has been provided to indicate the contribution of each feature to the model's prediction performance.

2. Related Works

In the realm of energy consumption prediction, a range of innovative methodologies have been explored. The works explored encompass hybrid models, advanced deep learning architectures, and ensemble techniques, each contributing novel approaches to accurate load forecasting. With applications ranging from residential to industrial sectors, these studies collectively offer insights into enhancing energy efficiency and management through innovative predictive tools. Kao et al. [11] developed a hybrid forecasting model to overcome the challenge of accurately predicting energy consumption, which is attributed to the non-linear nature of electricity consumption time series. The framework combines individual forecasting models such as ARIMA- genetic algorithm- support vector regression using the ensemble approach. The viability of the proposed framework was validated through a study using Taiwanese energy consumption data. Moreover, the framework outperformed other forecasting methods in terms of accuracy.

In their study, Ridwana et al. [12] highlight the need for energy-efficient building systems to address the significant energy usage in the expanding building sector. The ANN based model that incorporates data classification to enhance the precision of forecasting energy consumption on an hourly or sub-hourly basis for four buildings. The proposed models exhibit improved performance in assessing electricity demand compared to traditional regression models. This approach holds potential for applications in building energy conservation.

Abdullatif Baba [13] examines the performance of three methods for forecasting daily power usage in an industrial area. The methods tested are a probabilistic approach based on Multiple Model Particle Filter, two ANNs with different numbers of hidden layers, and an adaptive ANN that adjusts its structure based on historical data. The study emphasizes the capabilities of artificial intelligence (AI) techniques, as demonstrated by a supplementary analysis that utilizes a genetic algorithm to propose an optimal generator outage schedule.

Ngoc-Son Truong and colleagues [14] proposed utilizing additive ANNs (AANNs) to estimate the energy usage in residential buildings using a dataset obtained from a building that had a solar PV system. Their AANNs model displayed better accuracy, with MAPE of 14.04% and a MAE of 111.98 Watt-hour. The AANNs model outperformed support vector regression (SVR) by 103.75% in MAPE and traditional ANNs by 4.6% in MAPE. The researchers concluded that among the tested models, the AANNs model demonstrated superior effectiveness in predicting energy consumption. They also noted that this model could serve as a valuable tool for building managers aiming to enhance energy efficiency.

The authors, Jui-Sheng Chou and Duc-Son Tran [15], conducted a review of machine learning methods that utilize real-time data to forecast energy usage in buildings. They assessed the performance of single, ensemble, and hybrid models and concluded that the hybrid model, which utilizes both forecasting and optimization techniques, displayed the highest accuracy. Their primary goal was to provide a comprehensive overview of short-term load forecasting techniques and support users in energy management planning.

Dorado Rueda et al. [16] suggests a deep learning architecture based on WaveNet to forecast load demand in France 24 hours ahead. It shows superior performance compared to traditional statistical approaches such as ARIMA and other deep learning models. The study concludes that the proposed architecture can provide accurate and robust load demand forecasting in complex energy systems.

Ramos and colleagues [17] proposed a new methodology for energy consumption forecasting that utilizes an ANN and utilizing progressive learning techniques. The ANN undergoes daily training to ensure the forecasting model remains up-to-date. The study utilized a dataset covering a period of 16 months, with data collected at 5-minute intervals from an actual industrial facility. The research indicated that WAPE (Weighted Absolute Percentage Error) was a more dependable measure for forecast performance, and the proposed method maintained low forecast errors ranging from 8.5% to 13.5%. The ANN model with 128 neurons in the intermediate layers and a learning rate of 0.005 exhibited the highest precision. The outcomes indicate the advantages of the suggested method, including reduced energy consumption and optimized energy management, resulting in decreased energy costs.

Taheri, S et al. [18] suggests a hybrid prediction model for load forecasting in modern energy systems, where the uncertainty, non-linearity, and non-stationarity of signals pose a challenge. The proposed model integrates long short-term memory (LSTM) with empirical mode decomposition for California ISO dataset, resulting in improved accuracy in demand forecasting. The performance was compared to that of single LSTM, XGBoost, and logistic regression models. They demonstrated a significant improvement in MAPE for both short- and long-term prediction compared to the other models. Deep learning models may be preferred for more complex problems with large datasets and when the data exhibits spatial or temporal dependencies [19].

In our proposed work, we utilize the CatBoost algorithm, which is a fast prediction algorithm with a symmetric tree structure, for building the regression model. Additionally, we perform feature selection to remove irrelevant features and use GridSearchCV with 5-fold cross-validation to optimize hyperparameters.

3. Material and Methods

3.1. Data

The UCI steel industry energy consumption dataset [20] consists of various parameters related to energy consumption, which are listed in Table 1. This dataset is available in the publicly available database [20]. The dataset has 11 attributes and 35040 instances. The column 'date' is of type object, and the other columns are of either float64, int64, or object type. The dataset is complete, and there are no null or missing values present. The ‘Usage_kWh’ is the target variable. The dataset contains electricity consumption data for every 15 minutes in a year-long period.

Table 1. Attributes description

Attribute	Description
Usage_kWh	The continuous energy consumption in kilowatt-hour (kWh)
Lagging_Current_Reactive.Power_kVarh	Continuous kVarh for lagging current reactive power.
Leading_Current_Reactive_Power_kVarh	The continuous measurement of the leading current reactive power in kVarh.
CO2(tCO2)	Continuous ppm
Lagging_Current_Power_Factor	Power factor in %
Leading_Current_Power_Factor	Power factor in %
NSM	Continuous variable indicating the seconds elapsed since midnight
Week status	Categorical (Weekend (0) or a Weekday(1))
Day of week	Week days
Load Type	The load is categorized into three categories: light, medium, and maximum load

Fig. 1 shows the proposed energy consumption prediction model. The objective of this regression model is to predict the energy consumption ('Usage_kWh') based on the features listed in Table 1, excluding the 'NSM', 'WeekStatus', and 'Day_of_week' features. Table 2 provides the various statistics and characteristics for different features of the dataset. The mean is the average value of the attribute across all instances. For example, the mean energy consumption (Usage_kWh) across all instances is approximately 27.387 kilowatt-hours. Standard deviation measures the dispersion or spread of values around the mean. It gives you an idea of how much the values vary from the mean. A higher standard deviation indicates more variability. The minimum value observed for the attribute. For instance, the minimum energy consumption (Usage_kWh) recorded in the dataset is 0 kWh. The % indicates the first, second and third quartile. The maximum value observed for the attribute. For example, the maximum energy consumption (Usage_kWh) recorded in the dataset is 157.18 kWh. The dataset has the values every 15 minutes. This detailed temporal resolution challenges in handling and processing, requiring careful consideration during analysis. The inclusion of categorical variables such as 'Week Status' and 'Day of Week' introduces a need for appropriate encoding or handling during modeling, ensuring their meaningful incorporation into the predictive model.

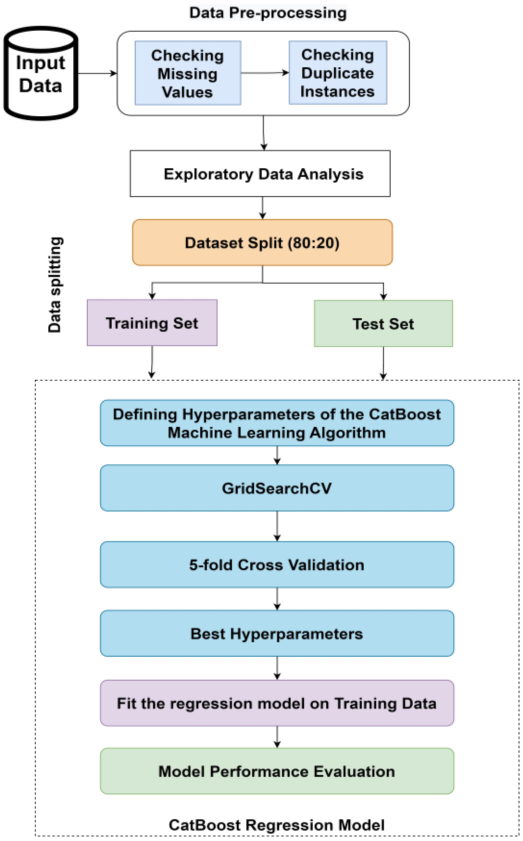
Data Pre-processing
Input Data
Checking Missing Values
Checking Duplicate Instances
Exploratory Data Analysis
Dataset Split (80:20)
Data splitting
Training Set
Test Set
Defining Hyperparameters of the CatBoost Machine Learning Algorithm
GridSearchCV
5-fold Cross Validation
Best Hyperparameters
Fit the regression model on Training Data
Model Performance Evaluation
CatBoost Regression Model

Fig. 1. Proposed energy consumption prediction model

Table 2. Dataset characteristics

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Usage_kWh	35040	27.387	33.444	0	3.2	4.57	51.2375	157.18
Lagging_Current_Reactive.Power_kVarh	35040	13.035	16.306	0	2.3	5	22.64	96.91
Leading_Current_Reactive_Power_kVarh	35040	3.871	7.424	0	0	0	2.09	27.76
CO2(tCO2)	35040	0.012	0.016	0	0	0	0.02	0.07
Lagging_Current_Power_Factor	35040	80.578	18.921	0	63.32	87.96	99.0225	100
Leading_Current_Power_Factor	35040	84.368	30.457	0	99.7	100	100	100
NSM	35040	42750	24940.53	0	21375	42750	64125	85500

3.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is to bring the most essential features of the data into focus for further analysis [21]. Fig. 2 displays the CO_2 Emission by weekday for different load types. The data suggests that CO_2 emissions are consistently low for the light load category (0.00 ppm), indicating minimal contribution to emissions. Conversely, the maximum load category consistently exhibits the highest emissions (0.03 ppm) on various weekdays. The medium load category falls in between, with emissions usually at 0.02 ppm, occasionally dropping to 0.01 ppm on certain days. The industry can focus on implementing load management strategies to reduce the reliance on maximum load operations, optimizing energy consumption, and distributing the load more efficiently to minimize CO_2 emissions associated with high load demands. Regardless of the load type, the industry can further reduce overall CO_2 emissions by implementing energy efficiency measures, including upgrading to energy-efficient technologies, optimizing processes, and conducting regular energy audits. In pursuit of further emission reduction, the industry can explore integrating renewable energy sources into its operations. By investing in solar, wind, or other clean energy solutions, the industry can supplement traditional energy sources, leading to reduced emissions and a more sustainable energy mix.

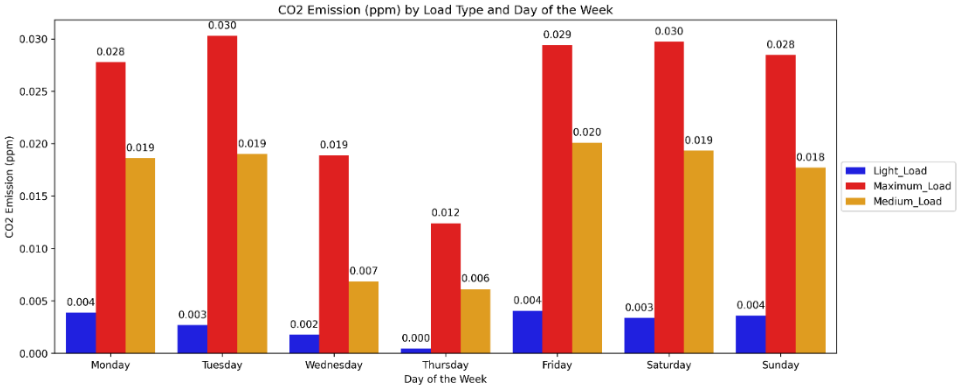

Fig. 2. CO_2 Emission by weekday for different load types

Fig. 3 demonstrates comparable effects, where it displays the lagging current reactive power for various load types categorized by weekday. Similarly, Fig. 4 illustrates the leading current reactive power by weekday for different load types. For medium loads, the leading current reactive power is maximum on Wednesday followed by Thursday. It is observed that the leading current reactive power is maximum for medium loads, while it is comparably low for maximum and light loads. On the other hand, lagging current reactive power is maximum for maximum loads, while it is low for light and medium loads. Lagging current reactive power continuous in kVARh is a measure of the total reactive power consumed by a load that has a lagging power factor over a given time period.

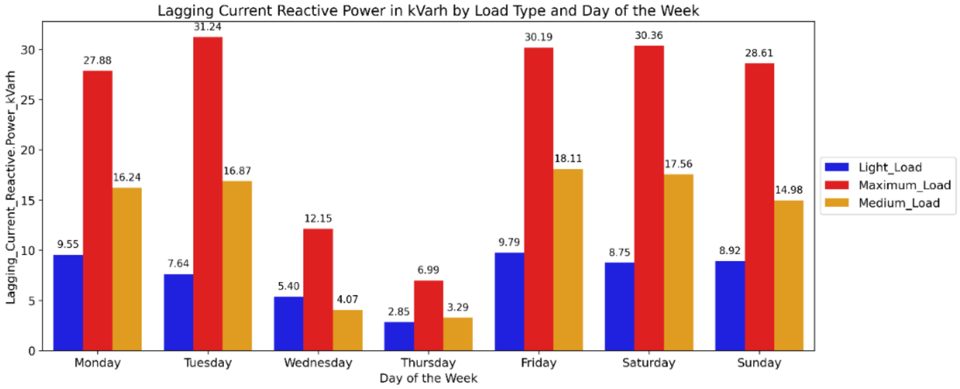

Fig. 3. Lagging current reactive power by weekday for different load types

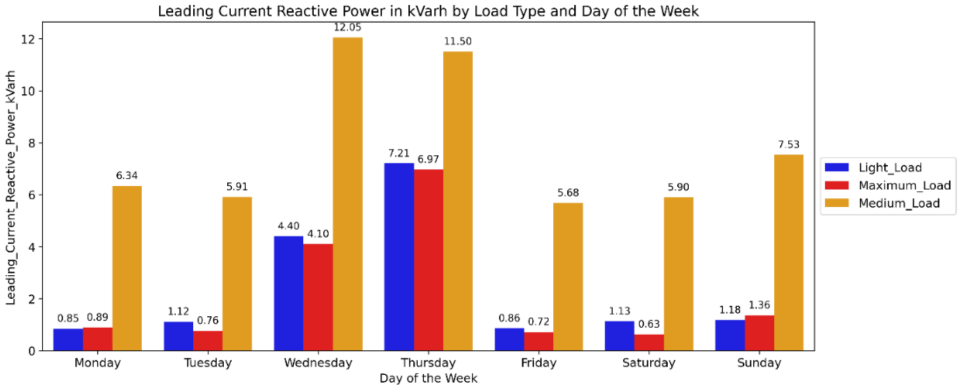

Fig. 4. Leading current reactive power by weekday for different load types

Reactive power is the power consumed by electrical devices to generate magnetic fields, which are required for the operation of devices like motors and transformers [22]. A lagging power factor occurs when the load consumes more reactive power than what is needed to meet its requirement of active power, causing the voltage and current to be out of phase. Continuous kVARh is a measure of the total reactive energy consumed by the load over time, and it is used to determine the compensation needed for reactive power to achieve the required power factor. Reactive power is the power in an AC circuit that does not contribute to the net power transferred to the load but instead alternately stores and returns energy to the source. It is measured in units of volt or kilovolt-amperes reactive (VAR or kVAR). The continuous kVarh is a unit of energy that indicates the accumulation of reactive power during a specific time frame and is widely utilized for measuring the overall reactive energy usage of a system or device.

Fig. 5 and Fig. 6 display the leading and lagging current power factors, respectively, by weekday for different load types. For maximum load conditions, the leading power factor is maximum in percentage, while for light and medium loads, the lagging power factor is comparably low. The lagging current power factor and leading current power factor columns represent the ratio of real power (kW) to apparent power (kVA), which indicates how efficiently the electrical power is being used. A lagging current power factor indicates that the current is lagging behind the voltage in the system, which can lead to higher energy consumption and lower efficiency [23].

Fig. 7 illustrates the energy consumption patterns specifically on weekdays. The data reveals that energy consumption attributed to light loads is higher on Fridays, Saturdays, Sundays, and Mondays. Conversely, energy consumption associated with light loads is comparatively lower on Thursdays. We can observe the same impact on CO2 emission, and lagging current power factor. Fig. 8 displays the energy consumption by day in the Steel Industry. It reveals that energy consumption is highest in February.

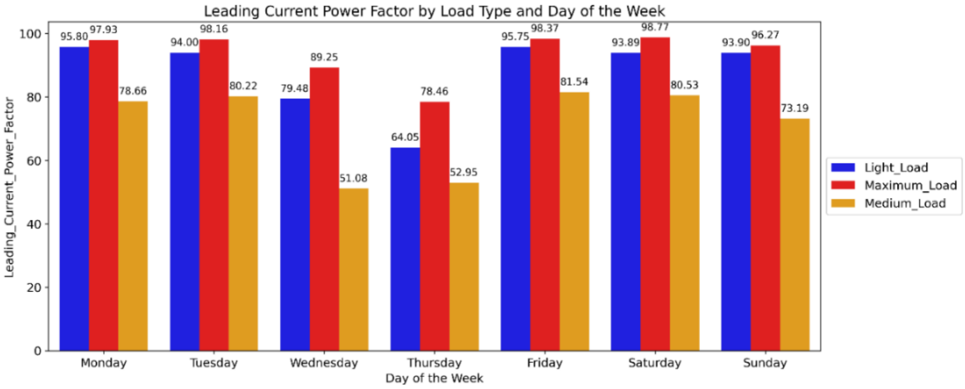

Fig. 5. Leading current power factor by weekday for different load types

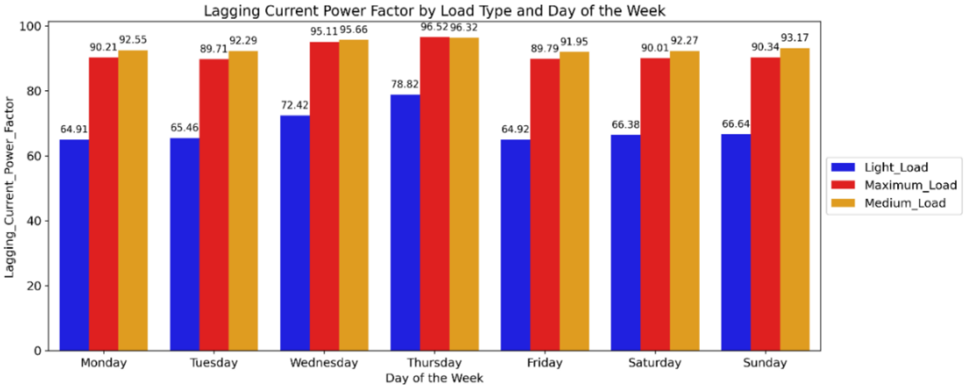

Fig. 6. Lagging current power factor by weekday for different load types

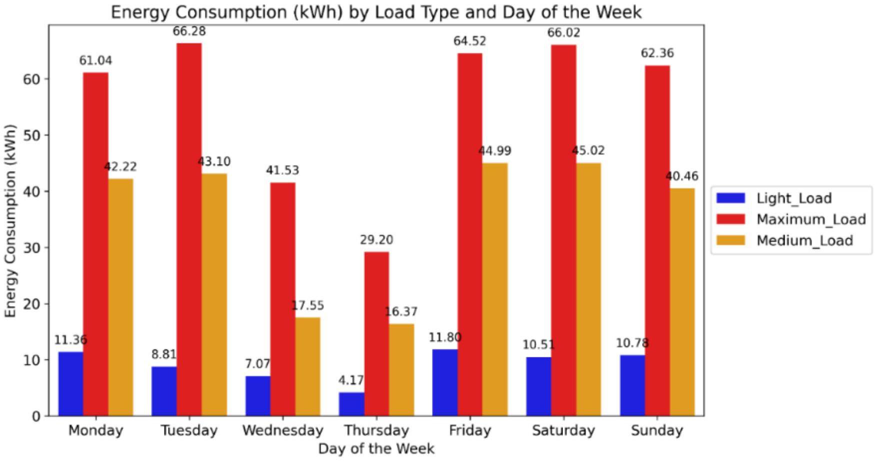

Fig. 7. Energy consumption by weekday for different load types

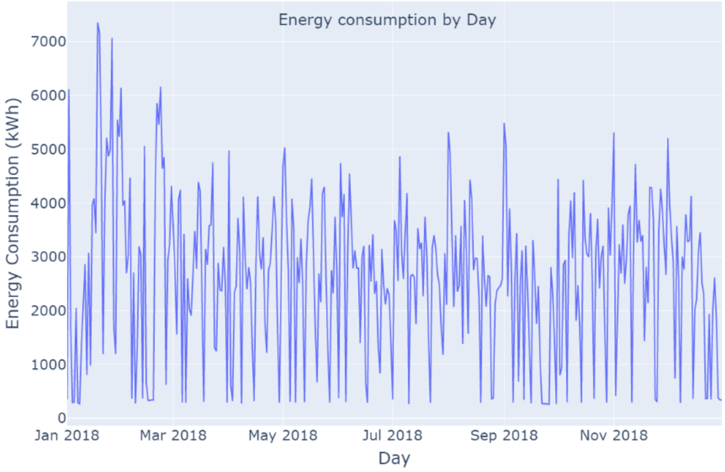

Fig. 8. Energy consumption by the day in steel industry

In Fig. 9, we see a heatmap of the dataset. Heatmaps provide a visualisation that is easy to comprehend by using different colours and sizes to represent data [24]. This heatmap displays the energy consumption in 'Usage_kWh' by 'Load_Type' and 'Day_of_week'. The rows represent the days of the week, while the columns represent the load types. The values in the cells indicate the energy consumption in kWh for that specific day of the week and load type combination. The colour scale shows the energy consumption level, with cool colours representing lower values and warm colours representing higher values. The annotations on the heatmap indicate the actual values for each cell, rounded to two decimal places. From the heatmap, it is observed that energy consumption varies based on the load type and the day of the week. Fridays and Thursdays tend to have the highest average energy consumption for both “Maximum Load” and “Medium Load,” while Sundays and Saturdays have the lowest consumption for these load types. Among the load types, “Maximum Load” generally exhibits the highest energy consumption, followed by “Medium Load” and then “Light Load.”

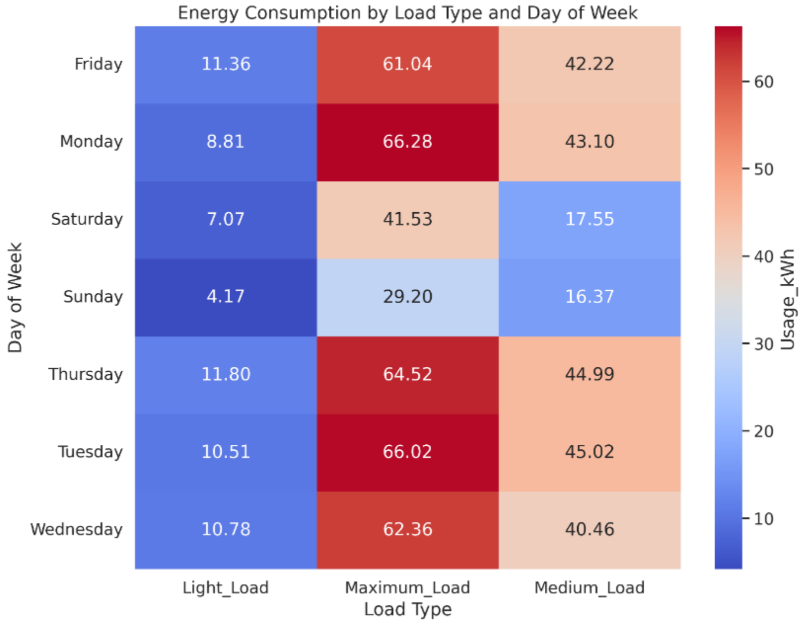

Fig. 9. Heatmap of the dataset

3.3. Data Pre-processing

The missing and duplicate values of the dataset have been checked, and label encoding has been performed on the categorical variables 'WeekStatus', 'Day_of_week', and 'Load_Type', converting them into integer variables. Label encoding is essential for ML/DL model development as it allows the model to interpret the data to make predictions or identify patterns. By encoding categorical variables, we transform them into a format easily used by machine learning algorithms [25].

The WeekStatus feature had two categories: “Weekday” and “Weekend”, which were encoded as 0 and 1, respectively. The Day_of_week feature had seven categories representing the days of the week, encoded as integers from 0 to 6, with 0 representing Monday and 6 representing Sunday. The Load_Type feature had three categories: “Light_Load”, “Medium_Load”, and “Maximum_Load”, encoded as integers 0, 1, and 2, respectively.

To characterize the presence of missing values, consider a matrix Y representing the complete dataset, which is divided into two components: Y_o representing the observed data, and Y_m representing the missing data [26]. Let R represent a matrix of missing values, defined as in equation (1).

$$R: = \begin{Bmatrix} 0\,; if\ Y\ is\ observed \\ 1\,; if\ Y\ is\ missing \end{Bmatrix} \tag{1}$$

3.4. Proposed CatBoost Machine Learning Algorithm

CatBoost is a boosting algorithm that is known for its fast prediction time and symmetric tree structure. It is approximately 8 times faster than XGBoost when it comes to prediction [27]. The CatBoost learning method is capable of automatically adjusting models to their environments. It can build complex connections between the output data and various types of incoming data, all of which are subject to change at any given time. This makes it possible to accurately predict the cost of power consumption.

Unlike other algorithms, CatBoost does not require any additional information to function well, such as domain knowledge on the impact of numerous features on energy consumption. Provided that sufficient training data is available, the learning process will identify all the relationships between the parameters automatically.

To implement CatBoost, various steps need to be taken, such as forming training and test subsamples for the algorithm, as well as a control subsample to evaluate the adequacy of the model. The control subsample ensures that the resulting model is satisfactory. Defining the input and output vectors is also crucial. The input vector for the model consists of the production system's input features that have an impact on energy consumption. The output vector, on the other hand, represents the predicted power consumption value for the given timeframe. Greedy Target-based Statistics [28] is defined as in equation (2).

$$\frac{\sum_{j=1}^{p}[X_{j,k} = X_{i,k}]\,Y_i}{\sum_{j=1}^{p}[X_{j,k} = X_{i,k}]} \tag{2}$$

Consider a provided dataset of observations $D = \{X_i, Y_i\}$ where i = 1, 2, ..., n. If we have a permutation $\sigma = (\sigma_1, \sigma_2, \sigma_3 \ldots, \sigma_n)$, the value $x\sigma_{p,k}$ is replaced with [28] and it is given in equation (3).

$$\frac{\sum_{j=1}^{p-1}[x\sigma_{j,k} = x\sigma_{p,k}]Y_i + aP}{\sum_{j=1}^{p-1}[x\sigma_{j,k} = x\sigma_{p,k}]Y_i + a} \tag{3}$$

4. Results and Discussions

4.1. Dataset Split

The dataset with 35,040 instances and 7 features has been considered for developing the energy consumption regression model. The dataset was segmented into training and testing sets using an 80:20 ratio. This approach ensures that the model has sufficient data to learn from while also being able to evaluate its performance on unseen data. This technique also helps prevent the model from overfitting and memorizing the training data, allowing it to perform well on new data [29].

4.2. Features Selection

The ‘date’, ‘NSM, ‘WeekStatus’ and 'Day_of_week' features has been discarded for developing the CatBoost regression model. 'date' is likely a unique identifier for each data point and does not provide any information that can improve the prediction accuracy [30].

The feature 'NSM' (Number of seconds from midnight) is considered redundant as other features such as 'Load_Type' already capture time-related information. Similarly, the feature 'WeekStatus' is a binary indicator for weekdays or weekends, which is not expected to greatly influence energy usage predictions. The categorical feature 'Day_of_week' indicating the specific day of the week is also not anticipated to have a substantial impact on energy consumption predictions. Therefore, these features are not relevant to the problem and have been discarded to simplify the model and improve its performance.

The other features listed in Table 1 have been selected because they are believed to have a strong relationship with the target variable, which is Usage_kWh. Specifically, Lagging Current_Reactive.Power_kVarh and Leading_Current_Reactive_Power_kVarh are measures of reactive power, which is a significant factor in estimating the total power consumption. CO2 (tCO2) is a measure of carbon dioxide emissions, which is directly related to energy consumption. Lagging_Current_Power_Factor and Leading_Current_Power_Factor are measures of power factor, which is also an important factor in determining energy consumption. Finally, Load_Type is a categorical variable that indicates the type of load, which could be useful in predicting energy consumption patterns for different types of loads.

4.3. GridSearchCV Optimization

The hyperparameter tuning using GridSearchCV with 5-fold cross validation has been performed. GridSearchCV is a technique in machine learning used for tuning hyperparameters. GridSearchCV is a technique that entails a comprehensive search across a predefined range of hyperparameters to find the best combination that maximizes the effectiveness of the CatBoost model [31].

In the context of the CatBoost regression model, GridSearchCV is employed to fine-tune crucial hyperparameters such as the number of trees (iterations), tree depth (depth), and learning rate (learning_rate) to optimize the model's performance. By tuning these hyperparameters, we can progress the model performance and achieve better results. Initially CatBoostRegressor model has the following hyperparameters:

Iterations: The number of trees included in the gradient boosting model. Three values are provided: 500, 1000, and 1500. depth: The maximum depth of each tree. Three values are provided: 4, 6, and 8. learning_rate: It defines the step size taken during each iteration of the training process. Three values are provided: 0.01, 0.1, and 0.5.

During the training process, the CatBoost model will iteratively improve its predictions on the training data by adjusting its parameters. The objective is to reduce the RMSE. After completing the training process, the model becomes capable of making predictions on fresh data.

The choice of k in k-fold cross-validation varies according to the size of the dataset and the computational resources available [32]-[36]. The value of k is set to 5 in this scenario, which results in the dataset being divided into 5 folds of the same size. The model undergoes training and testing five times, where each fold is utilized as the validation set once, while the remaining four folds are used for training purposes. This approach helps to assess the CatBoost model's effectiveness on various subsets of the data and reduces the likelihood of overfitting. The prediction error is estimated using cross-validation as follows:

$$CV(f) = \frac{1}{n}\sum_{i=1}^{n} T(Y_i, f^{-k(i)}(x_i)) \tag{4}$$

Here, k is the count of subsets, n represents the dataset's size, T is the loss function, and $f^{-k(i)}$ stands for the fitted function. GridSearchCV defines the hyperparameters to optimize as the number of iterations, depth of the tree, and learning rate. It then fits the GridSearchCV on the training data using these hyperparameters and uses negative mean squared error as the scoring metric.

After the hyperparameters have been identified, the code trains the CatBoost model with the best hyperparameters obtained from the GridSearchCV. It sets the number of iterations, depth, and learning rate of the model to the values identified by GridSearchCV, and then fits the model on the training data with verbose set to False. The hyperparameters that were obtained after performing GridSearchCV with 5-fold cross validation are: 'depth': 4, 'iterations': 1500, 'learning_rate': 0.5, Loss Function: RMSE and Random Seed: 42.

4.4. Energy Consumption Prediction

The residual plot of the CatBoost regression model is presented in Fig. 10, which is a useful tool for evaluating the effectiveness of the CatBoost regression model. The blue dots represent the model's predictive capability on the training set, while the red dots represent the performance on the test set. A desirable model should have residuals that are randomly distributed around the zero line, implying that the model is not systemically overestimating or underestimating the target variable.

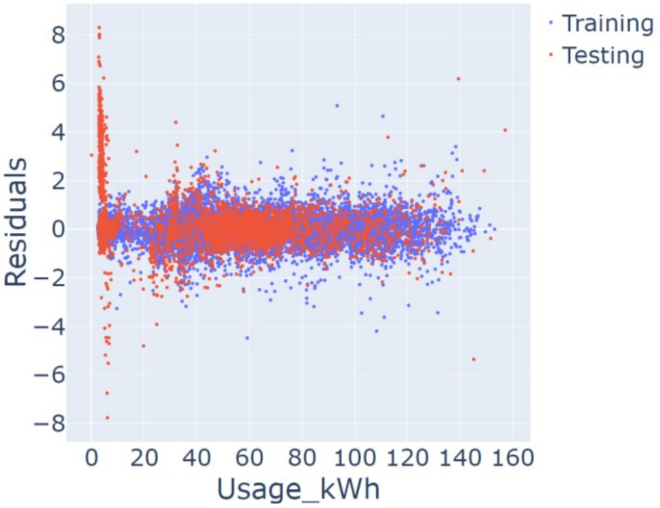

Fig. 10. Residual plot

4.5. Evaluation Matrices

There are several measurements for the performance of regression model predictions, but here we are using dour measurements of a given dataset [37]–[39]. RMSE: It represents the standard deviation of the prediction errors as shown in equation (5). It provides insight into the concentration of the data around the line of best fit. Equation (6) represents the coefficient of determination or R2, while equation (7) represents the MAPE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu)^2} \tag{6}$$

$$MAPE = \frac{1}{n} \times \sum_{1}^{n} \frac{|Actua - Predicted|}{Actual} \times 100 \tag{7}$$

where n is the total number of observations. MAPE is useful for measuring the accuracy of a model when the scale of the target variable varies widely.

RMSE is particularly relevant for energy consumption prediction models as it provides a comprehensive understanding of the prediction errors in the same unit as the target variable (e.g., kilowatt-hours). This makes it easy to interpret and directly assess the accuracy of the model's predictions. R2 is valuable for understanding the proportion of variability in energy consumption that the model explains. MAPE is suitable for energy consumption prediction models because it provides a percentage measure of the accuracy of predictions. This is particularly relevant in applications like energy forecasting, where stakeholders often need to understand the magnitude of errors relative to the actual consumption.

4.6. Model Performance Evaluation

Table 3 presents the evaluation metrics for the CatBoost ML algorithm on the training set and the test set. The training set has an Mean Squared Error (MSE) of 0.146, while the test set has an MSE of 1.152. The RMSE, which indicates the proximity of predicted values to actual values, is 0.382 for the training set and 1.073 for the test set. A higher R-squared value signifies a greater ability of the model to explain the variability in the target parameter, whereas a higher value represents best fit. The R-squared value is 0.999 for the training set and 0.998 for the test set. The MAPE is 1.139% for the training set and 1.142% for the test set. The CatBoost ML algorithm demonstrates strong performance in predicting energy consumption for the dataset.

Table 3. Evaluation matrices of CatBoost ML algorithm

ML Algorithm	MSE	RMSE	R2	MAPE
CatBoost – Training Set	0.146	0.382	0.999	1.139
CatBoost – Test Set	1.152	1.073	0.998	1.142

Fig. 11 shows the feature importance plot of the CatBoost regression model. The provided values represent the feature importance scores of different features in a predictive model. Feature importance scores indicate the contribution of each feature to the model's prediction performance. Specifically, the higher the importance score, the more influential the feature is in making accurate predictions. CO_2 (tCO2) feature has the highest importance score of 64.552. This indicates that the CO_2 emissions have the most significant impact on the model's predictions. Lagging_Current_Reactive.Power_kVarh feature has an importance score of 27.478. It is the second most important feature in the model. Reactive power consumption refers to the power oscillations between the source and the load. A high value of this feature might suggest inefficiencies in the system, affecting energy consumption. Lagging_Current_Power_Factor has an importance score of 4.584, this feature contributes to the model's predictions. Power factor is a measure of how effectively electrical power is being used in the system. A low power factor can indicate inefficient use of energy. 'Load_Type' is categorical feature has an importance score of 1.922. While not as significant as the previous features, it still contributes to the model's predictions. The type of load might affect energy consumption patterns. 'Leading_Current_Power_Factor' feature has an importance score of 1.283. Similar to lagging power factor, the leading power factor measures the efficiency of power usage. Its contribution is less compared to other features. 'Leading_Current_Reactive_Power_kVarh' feature has the lowest importance score of 0.181. It seems to have the least influence on the model's predictions compared to other features.

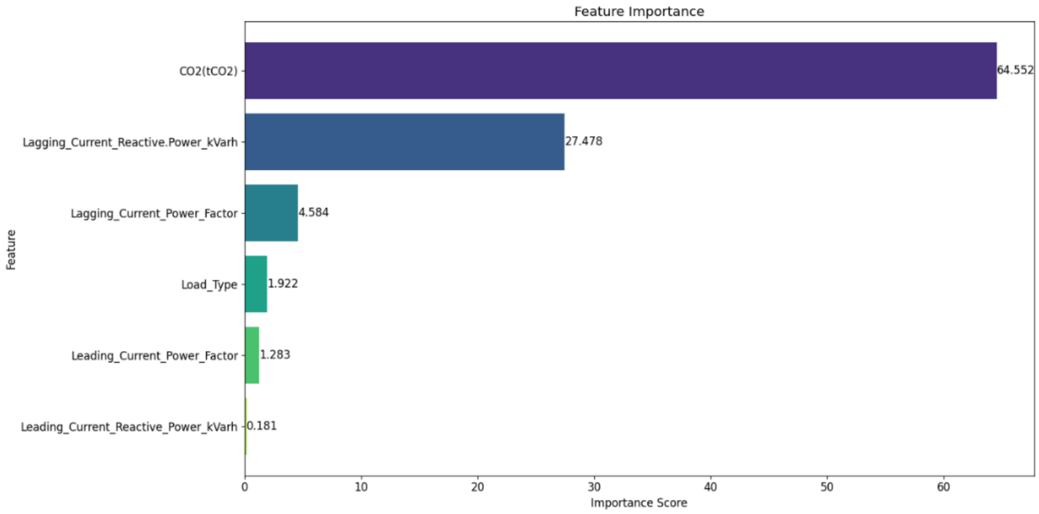

Fig. 11. Feature importance plot

Table 4 compares the proposed CatBoost approach with existing methods. The evaluation metrics used are RMSE and MAPE for both training and test sets. The comparison includes GLM, CART, SVM, kNN, and RF. The results clearly indicate that the proposed CatBoost model outperforms all the existing methods in terms of both RMSE and MAPE on the training set. Additionally, on the test set, CatBoost performs better than GLM, CART, and kNN in terms of RMSE and outperforms all the methods in terms of MAPE. These results strongly suggest that CatBoost is a promising approach for energy consumption prediction.

One of the reasons why CatBoost outperforms some of the other methods like SVM and kNN is its ability to efficiently handle both numerical and categorical features without losing valuable information. CatBoost's unique combination of gradient boosting and ordered boosting optimizes the training process, leading to faster convergence and improved performance. The technique of "ordered boosting" in CatBoost prevents overfitting, reducing the risk of memorizing the training data and ensuring better generalization to unseen data, addressing a common pitfall in other boosting algorithms.

Table 4. Comparison with existing approach

Method	**ML Algorithm**	**Training Set**		**Test Set**	
		RMSE	MAPE	RMSE	MAPE
Sathishkumar et al. [40]	GLM	4.61	16.01	4.85	16.13
	CART	3.27	13.88	3.46	14.10
	SVM	1.89	4.67	1.97	4.87
	KNN	1.59	2.89	2.99	5.33
	RF	0.5	2.45	1.12	1.28
Proposed Approach	CatBoost	0.382	1.139	1.073	1.142

Furthermore, CatBoost incorporates built-in regularization techniques that help control model complexity and improve generalization. This is essential for achieving good performance on test data and avoiding the issue of overfitting. Additionally, CatBoost's robustness to handle categorical variables and missing data further contributes to its superior performance compared to traditional algorithms like GLM, CART, SVM, kNN, and RF in this specific energy consumption prediction task.

The evaluation results and the analysis clearly show that CatBoost is a powerful and effective approach for accurately predicting energy consumption patterns, making it a valuable tool for energy management and conservation. Its ability to handle diverse features and its optimization techniques make it a promising choice for real-world applications. By utilizing CatBoost, organizations can gain valuable insights into energy usage, optimize resource allocation, and make informed decisions to reduce costs and promote sustainability.

5. Conclusion

In this study, we developed a CatBoost regression model to accurately predict energy consumption patterns based on various parameters related to energy usage. The dataset from the UCI steel industry energy consumption dataset provided valuable insights into electricity consumption every 15 minutes throughout the year.

The CatBoost algorithm demonstrated exceptional performance in predicting energy consumption, outperforming existing methods such as GLM, CART, SVM, kNN, and RF in terms of both RMSE and MAPE on both the training and test sets. The high-performance metrics of CatBoost validate the robustness and effectiveness of this approach. The elimination of irrelevant features has led to improved model efficiency and enhanced generalization capabilities.

Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. Overall, the developed CatBoost regression model with optimized hyperparameters can be utilized to accurately predict energy consumption patterns, making it a valuable tool for energy management and conservation.

However, it is important to note that the model's predictions may be influenced by factors not included in the dataset, such as weather conditions or seasonal trends. In the future, further research can explore the integration of external data sources to enhance the model's accuracy, such as weather data or market trends. Additionally, implementing a web-based dashboard or visualization tool can provide real-time insights and alerts for energy consumption patterns, enabling proactive maintenance and optimizing energy costs. As industries increasingly prioritize sustainable practices, the methodologies and models developed herein could serve as a cornerstone for the implementation of more sophisticated and efficient energy management systems, fostering a new era of innovation and conservation in the realm of industrial energy consumption.

Data Availability Statement: The dataset is available in publicly accessible database.

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

Funding: This research received no external funding.

References

[1] W. Long *et al.*, "Quantitative assessment of energy conservation potential and environmental benefits of an iron and steel plant in China," *Journal of Cleaner Production*, vol. 273, p. 123163, 2020, https://doi.org/10.1016/j.jclepro.2020.123163.

[2] B. Gajdzik, R. Wolniak, and W. W. Grebski, "Electricity and Heat Demand in Steel Industry Technological Processes in Industry 4.0 Conditions," *Energies*, vol. 16, no. 2, p. 787, 2023, https://doi.org/10.3390/en16020787.

[3] B. Sizirici, Y. Fseha, C. S. Cho, I. Yildiz, and Y. J. Byon, "A Review of Carbon Footprint Reduction in. Construction Industry, from Design to Operation," *Materials* vol. 14, no. 20, p. 6094, 2021, https://doi.org/10.3390/ma14206094.

[4] B. A. Salih, P. Wongthongtham, G. Morrison, K. Coutinho, M. Al-Okaily, and A. Huneiti, "Short-term renewable energy consumption and generation forecasting: A case study of Western Australia," *Heliyon*, vol. 8, no. 3, p. e09152, 2022, https://doi.org/10.1016/j.heliyon.2022.e09152.

[5] M. Awad and R. Khanna, "Machine Learning," *Efficient Learning Machines*, pp. 1-8, 2015, https://doi.org/10.1007/978-1-4302-5990-9_1.

[6] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Developments in the Built Environment*, vol. 5, p. 100037, 2021, https://doi.org/10.1016/j.dibe.2020.100037.

[7] M. R. Braun, H. Altan, and S. B. M. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," *Applied Energy*, vol. 130, pp. 305-313, 2014, https://doi.org/10.1016/j.apenergy.2014.05.062.

[8] J. Fattah, L. Ezzine, Z. Aman, H. E. Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, vol. 10, 2018, https://doi.org/10.1177/1847979018808673.

[9] R. Tehseen, M. S. Farooq, and A. Abid, "Earthquake Prediction Using Expert Systems: A Systematic Mapping Study," *Sustainability*, vol. 12, no. 6, p. 2420, 2020, https://doi.org/10.3390/su12062420.

[10] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, p. 83, 2019, https://doi.org/10.1038/s41524-019-0221-0.

[11] Y. S. Kao, K. Nawata, and C. Y. Huang, "Predicting Primary Energy Consumption Using Hybrid ARIMA and GA-SVR Based on EEMD Decomposition," *Mathematics*, vol. 8, no. 10, p. 1722, 2020, https://doi.org/10.3390/math8101722.

[12] I. Ridwana, N. Nassif, and W. Choi, "Modeling of Building Energy Consumption by Integrating Regression Analysis and Artificial Neural Network with Data Classification," *Buildings*, vol. 10, no. 11, p. 198, 2020, https://doi.org/10.3390/buildings10110198.

[13] A. Baba, "Advanced AI-based techniques to predict daily energy consumption: A case study," *Expert Systems with Applications*, vol. 184, p. 115508, 2021, https://doi.org/10.1016/j.eswa.2021.115508.

[14] N. S. Truong, N. T. Ngo, and A. D. Pham, "Forecasting Time-Series Energy Data in Buildings Using an Additive Artificial Intelligence Model for Improving Energy Efficiency," *Computational Intelligence and Neuroscience*, vol. 2021, 2021, https://doi.org/10.1155/2021/6028573.

[15] J. S. Chou and D. S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," *Energy*, vol. 165, pp. 709-726, 2018, https://doi.org/10.1016/j.energy.2018.09.144.

[16] F. D. Rueda, J. D. Suárez, and A. D. R. Torres, "Short-Term Load Forecasting Using Encoder-Decoder WaveNet: Application to the French Grid," *Energies*, vol. 14, no. 9, p. 2524, 2021, https://doi.org/10.3390/en14092524.

[17] D. Ramos, P. Faria, Z. Vale, J. Mourinho, and R. Correia, "Industrial Facility Electricity Consumption Forecast Using Artificial Neural Networks and Incremental Learning," *Energies*, vol. 13, no. 18, p. 4774, 2020, https://doi.org/10.3390/en13184774.

[18] S. Taheri, B. Talebjedi, and T. Laukkanen, "Electricity Demand Time Series Forecasting Based on Empirical Mode Decomposition and Long Short-Term Memory," *Energy Engineering*, vol. 118, no. 6, pp. 1577–1594, 2021, https://doi.org/10.32604/EE.2021.017795.

[19] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022, https://doi.org/10.1016/j.inffus.2021.11.011.

[20] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-6, 2018, https://doi.org/10.1109/ICCUBEA.2018.8697857.

[21] C. Chatfield, "Exploratory data analysis," *European Journal of Operational Research*, vol. 23, no. 1, pp. 5-13, 1986, https://doi.org/10.1016/0377-2217(86)90209-2.

[22] D. Błaszczok, T. Trawiński, M. Szczygieł, and M. Rybarz, "Forecasting of Reactive Power Consumption with the Use of Artificial Neural Networks," *Electronics* vol. 11, no. 13, 2022, https://doi.org/10.3390/electronics11132005.

[23] K. Sekar, K. Kanagarathinam, S. Subramanian, E. Venugopal, and C. Udayakumar, "An Improved Power Quality Disturbance Detection Using Deep Learning Approach," *Mathematical Problems in Engineering*, vol. 2022, 2022, https://doi.org/10.1155/2022/7020979.

[24] B. C. B. Haarman, R. F. R. D. Lek, W. A. Nolen, R. Mendes, H. A. Drexhage, and H. Burger, "Feature-expression heat maps – A new visual method to explore complex associations between two variable sets," *Journal of Biomedical Informatics*, vol. 53, pp. 156-161, 2015, https://doi.org/10.1016/j.jbi.2014.10.003.

[25] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang B, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data," *Frontiers in Energy Research*, vol. 9, 2021, https://doi.org/10.3389/fenrg.2021.652801.

[26] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, https://doi.org/10.1093/biomet/63.3.581.

[27] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of Big Data*, vol. 7, no. 94, 2020, https://doi.org/10.1186/s40537-020-00369-8.

[28] Y. Zhang, Z. Zhao, and J. Zheng, "CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China," *Journal of Hydrology*, vol. 588, p. 125087, 2020, https://doi.org/10.1016/j.jhydrol.2020.125087.

[29] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166-176, 2020, https://doi.org/10.1080/00401706.2021.1921037.

[30] K. Kanagarathinam, D. Sankaran, and R. Manikandan, "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset," *Data & Knowledge* Engineering, vol. 140, p. 102042, 2022, https://doi.org/10.1016/j.datak.2022.102042.

[31] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, Bernardo, L. F. A. Bernardo, and H. M. Hussein, “Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques,” *Materials*, vol. 15, no. 21, p. 7432, 2022, https://doi.org/10.3390/ma15217432.

[32] S. Prusty, S. Patnaik, and S. K. Dash, “SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer,” *Frontiers in Nanotechnology*, vol. 4, 2022, https://doi.org/10.3389/fnano.2022.972421.

[33] S. A. Agnes, A. A. Solomon, and K. Karthick, “Wavelet U-Net++ for accurate lung nodule segmentation in CT scans: Improving early detection and diagnosis of lung cancer,” *Biomedical Signal Processing and Control*, vol. 87, p. 105509, 2024, https://doi.org/10.1016/j.bspc.2023.105509.

[34] K. Kanagarathinam, S. K. Aruna, S. Ravivarman, M. Safran, S. Alfarhood, and W. Alrajhi, “Enhancing Sustainable Urban Energy Management through Short-Term Wind Power Forecasting Using LSTM Neural Network,” *Sustainability*, vol. 15, no. 18, p. 13424, 2023, https://doi.org/10.3390/su151813424.

[35] G. Ponkumar, S. Jayaprakash, and K. Kanagarathinam, “Advanced Machine Learning Techniques for Accurate Very-Short-Term Wind Power Forecasting in Wind Energy Systems Using Historical Data Analysis,” *Energies*, vol. 16, no. 14, p. 5459, 2023, https://doi.org/10.3390/en16145459.

[36] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, “Air quality prediction by machine learning models: A predictive study on the indian coastal city of Vishakhapatnam,” *Chemosphere*, vol. 338, p. 139518, 2023, https://doi.org/10.1016/j.chemosphere.2023.139518.

[37] G. Li *et al*., “Performance of Regression Models as a Function of Experiment Noise,” *Bioinformatics and Biology Insights*, vol. 15, 2021, https://doi.org/10.1177/11779322211020315.

[38] K. Sekar, S. Kumar. S, and K. Karthick, “Power Quality Disturbance Detection using Machine Learning Algorithm,” *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*, pp. 1-5, 2020, https://doi.org/10.1109/ICADEE51157.2020.9368939.

[39] B. S. Kumar, R. Cristin, K. Karthick, and T. Daniya, “Study of Shadow and Reflection based Image Forgery Detection,” *2019 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-5, 2019, https://doi.org/10.1109/ICCCI.2019.8822057.

[40] V. E Sathishkumar *et al*., “Industry Energy Consumption Prediction Using Data Mining Techniques,” *International Journal of Energy*, vol. 11, no. 1, pp. 7-14, 2020, http://dx.doi.org/10.21742/ijeic.2020.11.1.02.