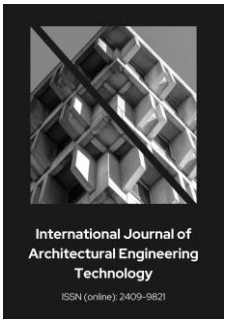
Published by Avanti Publishers

International Journal of Architectural Engineering Technology

ISSN (online): 2409-9821



Predicting Building Primary Energy Use Based on Machine Learning: Evidence from Portland

Yin Junjia*, Aidi Hizami Alias, Nuzul Azam Haron and Nabilah Abu Bakar

Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, Selangor, Malaysia

ARTICLE INFO

Article Type: Research Article

Academic Editor: Abdelhakim Mesloub

Keywords:
Primary energy use
Sustainability optimization
Building energy prediction
Machine learning algorithms

Timeline:
Received: November 03, 2024
Accepted: December 21, 2024
Published: December 28, 2024

Citation: Junjia Y, Alias AH, Haron NA, Abu-Bakar N. Predicting building primary energy use based on machine learning: Evidence from Portland. Int J Archit Eng Technol. 2024; 11: 124-139.

DOI: https://doi.org/10.15377/2409-9821.2024.11.7

ABSTRACT

Accurately predicting equivalent primary energy use (EPEU) in buildings is crucial for advancing energy-efficient design, optimizing operational strategies, and achieving sustainability goals in the built environment. This study aims to develop reliable prediction models for EPEU by leveraging a comprehensive and high-quality dataset from buildings in Portland, USA. To achieve this, a systematic machine learning framework is adopted, encompassing feature selection, data preprocessing, model training, and performance evaluation. Several state-of-the-art machine learning algorithms are applied, including Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Back-Propagation Neural Networks (BP). These models are trained using key features such as building type, gross floor area, construction year, and various operational characteristics that are known to significantly influence energy consumption patterns. The dataset is carefully cleaned and normalized to ensure model generalizability and minimize bias. Model performance is assessed using standard statistical metrics, including the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Among the tested models, ensemble learning methods—particularly RF and GBDT—consistently outperform others in terms of prediction accuracy, robustness, and stability across different building types. The results of this study not only highlight the potential of machine learning in energy prediction tasks but also provide actionable insights for architects, engineers, facility managers, and policymakers. By identifying the most influential variables and employing effective predictive models, this research supports data-driven decision-making processes aimed at improving building energy performance. Ultimately, the findings contribute to broader efforts in reducing carbon emissions and facilitating the transition toward more sustainable and energy-resilient urban environments.

*Corresponding Author
Email: gs64764@student.upm.edu.my
Tel: +(86) 18908453651

1. Introduction

The building sector is one of the largest energy consumers worldwide, accounting for a significant share of global primary energy use and carbon emissions. In the U.S., buildings account for 40% of total energy demand [1]. Accurate prediction of equivalent primary energy use (EPEU) is essential for optimizing building energy performance, informing policy decisions, and achieving sustainability goals. Fig. (**1**) illustrates the enormous consumption of primary energy in the U.S.'s residential, commercial, and industrial building sectors [2] from 1949 to 2023. The difference between primary and secondary energy sources lies in the form in which they exist in nature and the uses to which they are put [3]. Primary energy is unprocessed energy in nature. Examples include oil, natural gas, coal, solar energy, and wind energy [4]. Secondary energy is energy obtained by processing or converting primary energy. Examples include electricity, gasoline, diesel fuel, and hydrogen [5]. Primary energy sources are the basis of all energy conversion processes, while secondary energy sources are usually converted or processed for storage, transportation, and use. The primary sources of energy consumption in buildings are heating, ventilation, and air conditioning (HVAC) systems, water heating equipment, elevators, and lighting systems [6]. Traditional energy modeling approaches rely on deterministic methods, which may struggle to capture complex, nonlinear interactions among building characteristics, usage patterns, and environmental factors. Some data science methods are used in energy efficiency applications, such as optimization, neural networks, statistical analysis, and energy simulation [7]. However, a few studies have focused on comparing these algorithms' performance in building energy efficiency, especially primary energy prediction.

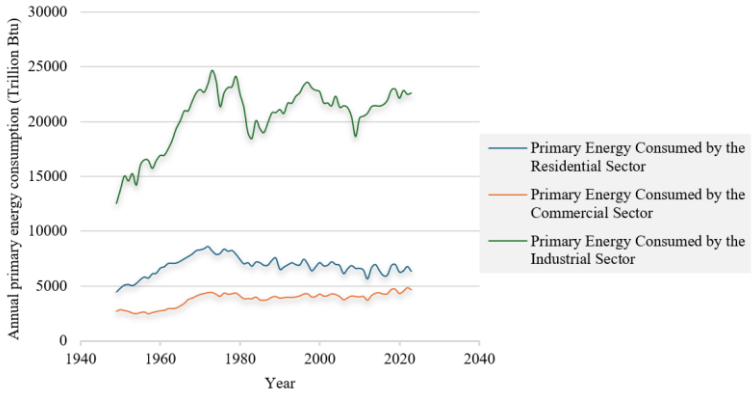


Figure 1: Energy consumption: residential, commercial, and industrial sectors.

Machine Learning is one of today's fastest-growing technological areas, at the intersection of computer science and statistics, and is at the heart of artificial intelligence and data science [8]. It enables the discovery of hidden patterns and improves predictive accuracy by utilizing large data sets [9]. However, a knowledge gap exists in applying and comparing multiple ML algorithms for building EPEU prediction in specific urban environments [10]. In contrast, the Portland region of the United States, with its diverse building stock and data on different energy consumption profiles, provides an ideal case study for exploring the potential of ML-based prediction methods.

The contributions of this study are: (1) Developing a more convenient primary energy forecasting model by using EPEU as a target variable. (2) Utilizing a comprehensive dataset of the Portland area to provide area-specific insights that are not adequately captured in the literature. (3) Systematically compare multiple ML algorithms, such as the Random Forest (RF) and CatBoost algorithms, to assess their performance and applicability in EPEU forecasting. In summary, this study contributes to the growing knowledge base on the application of ML in building energy consumption prediction. It provides actionable insights for optimizing energy use in urban environments.

2. Literature Review

The accurate prediction of energy use has been a critical area of research for decades, with traditional approaches primarily relying on physics-based models or statistical regression methods [11]. While these methods

provide a foundational understanding, they often require extensive domain expertise. They may lack the flexibility to account for the nonlinear interactions among factors such as building design, operational patterns, and climatic conditions [12].

In recent years, machine learning (ML) techniques have emerged as powerful tools for addressing these challenges. Studies have demonstrated the effectiveness of ML algorithms, including Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and Backward Propagation Neural Networks (BP), in capturing complex relationships and improving prediction accuracy [13]. However, most existing research focuses on either single algorithms or generic energy use intensity (EUI), with limited attention given to equivalent primary energy use (EPEU), which is a more comprehensive measure reflecting both direct and indirect energy consumption [14].

Moreover, many studies are based on datasets from specific regions or building types, leading to limited generalizability of findings [15]. With its diverse building stock and unique energy consumption patterns, the Portland area remains underexplored in ML-based EPEU prediction. Additionally, there is a lack of systematic comparison of ML algorithms to identify their relative strengths and weaknesses under different scenarios [16].

3. Materials and Methods

3.1. Materials

The Building Performance Database (BPD) is the largest dataset of information on the energy-related characteristics of commercial and residential buildings in the U.S. The BPD integrates, cleans, and anonymizes data collected by federal, state, and local governments, utilities, energy efficiency programs, building owners, and private companies and makes it available to the public. It can be accessed at https://bpd.lbl.gov/explore. The site allows users to explore the real estate industry and regional data and compare various physical and operational characteristics to understand market conditions and energy performance trends better. Notably, BPD is sponsored by the U.S. Department of Energy's Office of Building Technologies. Lawrence Berkeley National Laboratory and Earth Advantage developed the web application. The data for this study then comes from the 2019 Portland Building Energy Efficiency Report. This dataset includes individual building records publicly available in the Portland Building Energy Efficiency Report. The dataset has been cleaned and formatted according to the Portland Building Energy Efficiency Report rules and contains only a portion of the original data. In addition, building density and climate conditions in Portland are shown in Fig. (**2**). Portland's building density is moderate, and the downtown area is dominated by low-rise and mid-rise buildings that emphasize the integration of green space and community space. The climate is temperate and maritime, with mild, rainy winters and warm, dry summers, making it ideal for living and outdoor activities.

Table 1: Sample descriptive analysis (N=485).

ID	N	Minimum	Maximum	Mean	Standard Deviation
FT	485	1.00	10.00	6.6990	2.61751
FA	485	14810	3258912	116798.14	234278.672
YB	485	1880	2019	1966.29	35.716
SE	485	9	562	86.95	63.823
SoE	485	24	1188	186.08	130.690

Table **1** demonstrates the basic statistical characteristics of the 485 sample buildings, including floor area (FA), year of construction (YB), site energy intensity (SE), and source energy intensity (SoE). The FA ranged from 14,810 to 3,258,912 square feet, with a mean value of 116,798.14 and a large standard deviation, indicating that building sizes varied significantly in the sample. The building years spanned a wide range, with an average year of

construction of 1966. The mean values of SE and SoE were 86.95 and 186.08 kBtu/ft², respectively, both with high standard deviations, suggesting that the level of building energy consumption fluctuates widely, providing a basis for further categorization or modeling analyses.

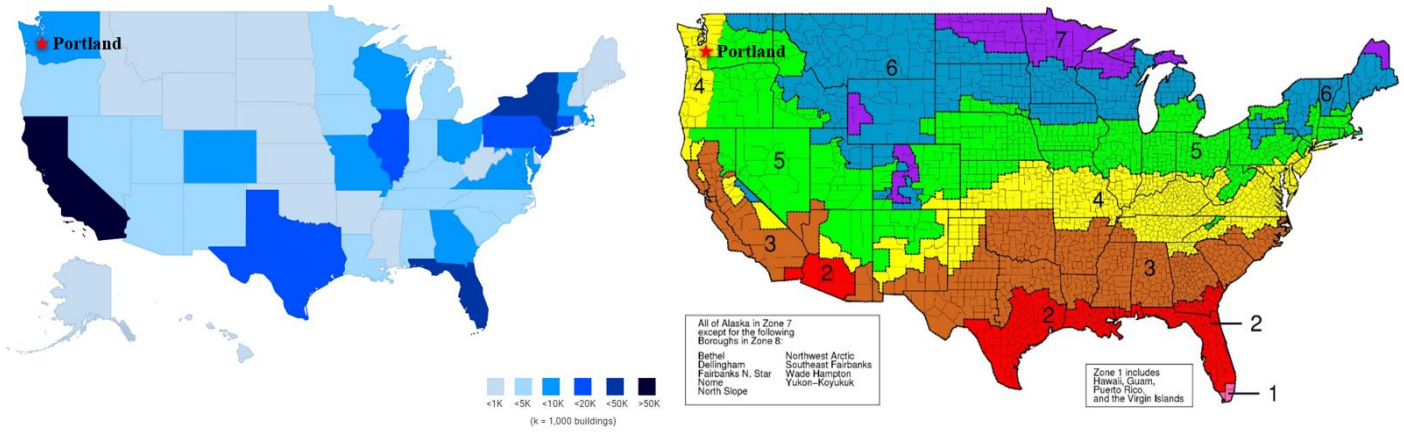


Figure 2: Building density and climate conditions in Portland.

3.2. Methods

As illustrated in Fig. (**3**), the methods used in this study are based on several well-defined steps that provide a solid and rigorous approach to achieving our goals. We are working on more advanced techniques for predicting primary energy consumption in buildings. The first step is data preprocessing, removing vacant entries in the source data and checking for information integrity. Step 2: Feature Extraction In this step, we manually extract feature items with complete information from the source data. This step can highlight patterns and essential information in the data. Next, we divide the data into two groups: 80% for training and 20% for the testing set. The third step is cross-validation learning. In this critical phase, we use algorithmic cross-validation to train our model. In this way, we can test different algorithms and select the best performance to regress the primary energy consumption of buildings, which is our specific task. The fourth step is testing and evaluation. Each algorithm is trained on preprocessed data and extracted features, and then the model is evaluated on an independent test dataset. We use a test dataset independent of the training and validation datasets to assess the model's performance.

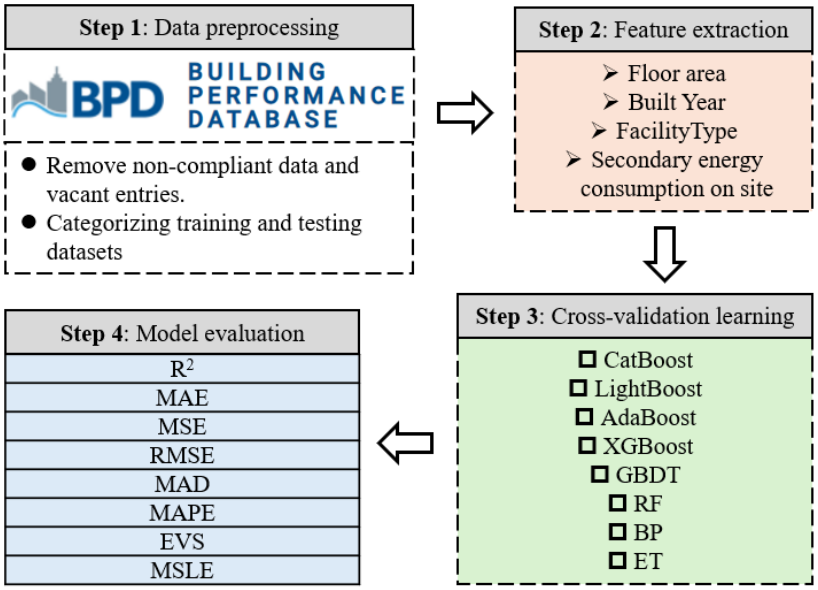


Figure 3: Research process.

Bayesian Optimization is used to regulate the hyperparameters of machine learning models [17]. Compared with traditional grid or random search, it effectively reduces the computational cost and improves efficiency by constructing a probabilistic model of the objective function and selecting the parameter combinations most likely to improve the model's performance at each step [18]. This approach is particularly suitable for modeling scenarios with large parameter space and high training costs and can find parameter configurations with better performance faster.

To ensure the robustness of the model evaluation, the study also adopts a 5-fold cross-validation strategy, which effectively reduces the evaluation bias that may be brought about by different data divisions [19]. In addition, to ensure the consistency of the distribution of categories between the training and testing sets, the study adopts stratified sampling to divide the data, to maintain a consistent proportion of each type of sample in each fold, avoiding the model's bias towards a certain type of samples, and improving the model's generalization ability and the reliability of the assessment results [20].

3.2.1. CatBoost Algorithm

CatBoost is a Gradient Boosting Decision Trees (GBDT) based algorithm optimized to handle categorical features and sequential data [21]. The steps of the algorithm are as follows: (1) Generate an initial model, starting with a simple model, usually the average of all target values; (2) Iteratively construct a tree, calculating the residuals (i.e., prediction error) of the current model, using the residuals to construct a new decision tree, fitting these residuals, and adding this tree to the model to reduce the error; (3) Update the model and repeat the iterations, gradually adding decision trees until a predetermined number of trees or other stopping conditions are reached [22]. The specific parameter settings are shown in Table **2**.

Table 2: CatBoost model parameter setting.

Parameter	Value
Training set ratio	0.8
Loss function	Logloss
Number of iterations	500
Learning rate	0.1
Maximum tree depth	6
Feature subset ratio	1.0
L2 regularization	3.0

3.2.2. LightGBM Algorithm

LightGBM is an efficient machine learning algorithm based on Gradient Boosting Decision Trees (GBDT), which is mainly used for learning tasks dealing with large-scale data and high-dimensional features [23]. It employs optimization strategies that enable faster training, lower memory footprint, and the ability to handle large amounts of data and features. The steps of the algorithm are as follows: (1) initialization, constructing an initial learner (tree) as the base model; (2) iterative training, constructing more learners sequentially by iterative means, each learner tries to correct the error of the previous learner; (3) gradient optimization, optimizing the model according to the gradient information in each iteration, to minimize the loss function of the model on the training set; (4) leaf node splitting, which selects the optimal features and splitting points according to the splitting gain and gradually generates a more complex decision tree structure; and (5) boosting learning, which improves the overall model's prediction ability by accumulating the prediction results of multiple simple models [24]. The specific parameter settings are shown in Table **3**.

Table 3: LightGBM model parameter setting.

Parameter	Value
Training set ratio	0.8
Booster type	but
Number of learners	100
Learning rate	0.1
Maximum tree depth	5
Maximum number of leaves in the tree	31
Minimum number of samples in child nodes	20
Minimum weight of child nodes	0.001
Minimum node splitting gain	0.0
Sample sampling rate	1.0
Single tree sampling rate	1.0
Sampling frequency	1

3.2.3. AdaBoost Algorithm

The core idea of the AdaBoost (Adaptive Boosting) algorithm is to combine multiple weak classifiers into one strong classifier [25]. The steps of the algorithm are as follows: (1) initialize weights, assign equal initial weights to each training sample; (2) train weak classifiers, train a weak classifier based on the weights of the current samples and calculate its error rate; (3) update weights, increase the weights of misclassified samples, so that the subsequent weak classifiers will pay more attention to these samples, and reduce the weight of the correctly classified samples; (4) combine classifiers, combine the weighted results of all the weak classifiers to form the final strong classifier; and (5) combine the weighted results of all the weak classifiers to form the final strong classifier. Classifiers combine the weighted results of all weak classifiers to form the final strong classifier; (5) AdaBoost can significantly improve classification performance by iterating the above steps repeatedly [26]. The specific parameter settings are shown in Table **4**.

Table 4: AdaBoost model parameter setting.

Parameter	Value
Training set ratio	0.8
Loss function	Linear
Number of learners	100
Learning rate	1.0

3.2.4. XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) is a gradient-boosting algorithm that builds a more potent model by iteratively training weak classifiers (usually decision trees) and integrating their predictions [27]. The key points include [28]: (1) XGBoost is a gradient-boosting algorithm that gradually improves the overall model performance by combining multiple weak learners. Each iteration corrects the error of the previous model round and progressively improves the overall model performance; (2) Decision tree-based learner, XGBoost uses decision trees as the base learner to form a powerful integrated model. Each decision tree splits the data by selecting the optimal split point based on the principle of gradient descent; (3) Regularization terms, it introduces regularization terms, including L1 regularization and L2 regularization, to control the complexity of the model and prevent overfitting, and the regularization terms control the complexity of the tree by introducing penalty terms in the loss

function; (4) XGBoost provides an intuitive way to evaluate the importance of features by analyzing the contribution of features at the split points of the tree. The specific parameter settings are shown in Table **5**.

Table 5: XGBoost model parameter setting.

Parameter	Value
Training set ratio	0.8
Lifter type	gbtree
Number of learners	100
Learning rate	0.1
Maximum tree depth	6
Sample sampling rate	1.0
Feature sampling rate	1.0
Smallest sub-node weight	1.0
Split gain threshold	0.0
L1 regularization	0.0
L2 regularization	1.0

3.2.5. GBDT Algorithm

GBDT (Gradient Boosting Decision Tree) is an integrated learning algorithm based on decision trees, which iteratively weighs multiple weak classifiers (decision trees) to improve the accuracy of the model [29]. In each iteration, GBDT calculates the residual (the difference between the actual value and the predicted value) based on the prediction result of the previous model round and uses the residual as the training target of the next round of the model. The steps of the algorithm are as follows [30]: (1) initialization, by fitting an initial model (e.g., the mean value) to get the initial predicted value; (2) calculation of residuals, by calculating the residuals (differences) between the predicted value and the actual value of the current model; (3) fitting the residuals, by fitting a regression tree (decision tree) to predict the residuals, to make the residuals decrease; (4) updating the model, by multiplying the prediction result of the regression tree by a learning rate (or step size) to update the current model; (5) Repeat iterations, repeating steps 2 to 4 until a preset number of iterations is reached or the residuals are already small enough; (6) Integrate the model, combining all of the regression trees to form the final integrated model, with the predictions of each tree weighted and summed together to obtain the final predicted values. The specific parameter settings are shown in Table **6**.

Table 6: GBDT model parameter setting.

Parameter	Value
Training set ratio	0.8
Loss function	Linear
Number of learners	100
Learning rate	0.1
Maximum tree depth	6
Sample sampling rate	1.0
Minimum number of samples for node splitting	2
Minimum number of leaf node samples	1
Model convergence parameters	0.001

3.2.6. BP Algorithm

BP Neural Network (Back-Propagation Network) consists of two processes: forward propagation of the signal and backpropagation of the error [31]. That is, the calculation of the error output is performed in the direction from input to output, while the adjustment of weights and thresholds is performed in the direction from production to input. In forward propagation, the input signal acts on the output node through the implicit layer and undergoes a nonlinear transformation to produce an output signal. If the actual output does not match the desired output, it is transferred to the process of backpropagation of the error. Error backpropagation is to back-propagate the output error through the hidden layer to the input layer by layer, apportion the error to all units in each layer, and use the error signal obtained from each layer as the basis for adjusting the weights of each unit [32]. By changing the strength of the connection between the input node and the hidden layer node and the strength of the connection between the hidden layer node and the output node, as well as the threshold value, the error decreases along the gradient direction. After repeated learning and training, the parameters of the network (weights and thresholds) corresponding to the minimum error are determined. The specific parameter settings are shown in Table **7**.

Table 7: BP model parameter setting.

Parameter	Value
Training set ratio	0.8
Hidden layer neuron settings	(100)
Activation function	ReLU
Weight optimization method	Adam
L2 regularization coefficient	1.0E-4
Initial learning rate	0.001
Learning rate optimization method	Constant
Minibatch size	Custom
Maximum number of iterations	200
Optimization tolerance	1.0E-4

3.2.7. RF Algorithm

Random Forest (RF) is an integrated algorithm, a classifier containing multiple decision trees. Compared with a single decision tree, the Random Forest algorithm will perform better and can effectively prevent the phenomenon of overfitting [33]. The steps of the construction process are as follows [34]: (1) Random Forest randomly selects n training samples each time there is a put back (which can be controlled by the parameters) to form a new training set; (2) Decision tree construction is performed with the newly selected samples, and when constructing the decision tree, not all the features are used but some of the features are used, and each time the split adopts a particular strategy to select one of them as the split attribute; (3) Repeat the second step of the process until it cannot be split anymore; (4) Repeat the second step of the process until it cannot be split anymore; (5) Repeat the second step of the process until it cannot be split anymore. process until it cannot be split again; (4) Repeat steps 1~3 to build multiple decision trees, and numerous decision trees form a random forest. Each decision tree produces a categorization result when predicting, and the category with the highest final vote is the final model prediction. The specific parameter settings are shown in Table **8**.

3.2.8. ET Algorithm

The Extremely Randomized Trees (ET) algorithm is an integrated learning method to improve the generalization ability and stability of the model by constructing multiple extremely randomized decision trees [35]. The main steps are as follows [36]: (1) Sample selection, randomly select samples from the training data; (2) Feature

Table 8: RF model parameter setting.

Parameter	Value
Training set ratio	0.8
Number of decision trees	100
Node split criteria	squared_error
Minimum number of samples for node splitting	2
Minimum number of leaf node samples	1
Maximum tree depth	No limit
Maximum number of features	Sqrt
Whether to put back sampling	Yes
Whether or not out-of-bag data testing	Yes

selection, at each node, randomly select a part of features; (3) Split point selection, for each selected feature, randomly select a split point; (4) Node splitting, use randomly selected features and split points to split the node to create child nodes; (5) Tree construction, repeat the steps above until the predefined stopping conditions are reached (e.g., maximum tree depth or minimum number of node samples); (6) Integration of results, which integrates the prediction results of all trees (using voting for classification problems and averaging for regression problems). This randomization gives the Extra Trees algorithm an advantage in handling high-dimensional data and preventing overfitting. The specific parameter settings are shown in Table **9**.

Table 9: Extra Trees model parameter setting.

Parameter	Value
Data preprocessing	None
Proportion of training set	0.8
Number of decision trees	100
Node split criteria	squared_error
Minimum number of samples for node splitting	2
Minimum number of leaf node samples	1
Maximum tree depth	No limit
Maximum number of features limit	Auto
Whether or not put back sampling	Yes
Whether out-of-bag data testing	Yes

4. Results and Discussion

4.1. Correlation Analysis

Correlation analysis is used to study the relationship between quantitative data, whether there is a relationship or not, and how close the relationship is [37]. As can be seen from Table **10** above, correlation analysis is used to study the correlation between Facility type (FT) and Floor area (FA), Year built (YB), Site Eui (SE), Source Eui (SoE), using Spearman's correlation coefficient to indicate the strength of the correlation. Spearman's correlation coefficients reveal that FT is negatively correlated with SE ($r = -0.257$, $p<0.01$) and SoE ($r = -0.215$, $p<0.01$), indicating that certain facility types tend to have lower energy use intensities. FA shows a weak positive correlation

with SoE (r = 0.090, p<0.05), suggesting that larger floor areas slightly increase source energy usage. YB is positively correlated with SE (r = 0.133, p<0.01) and SoE (r = 0.166, p<0.01), implying that buildings constructed more recently have slightly higher energy intensities. Notably, SE and SoE are strongly correlated (r = 0.944, p<0.01), reflecting their direct relationship as energy performance measures.

Table 10: Correlation analysis.

Items	Mean	S.D.	FT	FA	YBt	SE	SoE
FT	6.699	2.618	1				
FA	116798.140	234278.672	0.065	1			
YB	1966.291	35.716	-0.019	0.042	1		
SE	86.954	63.823	-0.257**	0.086	0.133**	1	
SoE	186.080	130.690	-0.215**	0.090*	0.166**	0.944**	1

* p<0.05 ** p<0.01.

4.2. Regression Analysis

The model evaluation result indicators are used to evaluate the advantages and disadvantages of the models and compare them. As shown in Table **11**, this study provides 8 evaluation indicators, among which 4 indicators, R2 value, MAE, MSE, and RMSE, are used more [38]. Second, more attention is usually paid to the evaluation results at the time of the test set. Third, if the metrics fit significantly better in the training set than in the test set, it implies an overfitting problem. Fourth, if the fit metrics are abnormal (not within the standard range), such as the R^2 value appearing to be less than 0, it means that the data fits the model poorly, and it is recommended to discard the model.

Table 11: Indicators for model evaluation.

Indicator	Description
R^2	The degree of fit indicator, the greater, the better between 0 and 1.
Mean Absolute Error (MAE)	L1 loss, the mean difference between true and fitted values; the closer to 0, the better.
Mean Square Error (MSE)	L2 loss, the mean sum of squared errors; the closer to 0, the better.
Root Mean Square Error (RMSE)	MSE open root sign, average gap value.
Median absolute error (MAD)	The absolute value of the residuals of the predicted value from the media, independent of outliers, the smaller, the better.
Mean Absolute Percentage Error (MAPE)	Mean percent error, independent of outliers, more minor is better.
Explainable Variance Score EVS	The measure of the model's strength in explaining data fluctuations, between [0,1] is that the more significant, the better.
Mean square logarithmic error (MSLE)	It penalizes underprediction more (less use) when RMSE is the same.

Table **12** summarizes the performance of eight machine learning models during training. CatBoost, GBDT, and XGBoost exhibit the highest R^2 values (close to 1), indicating excellent predictive accuracy, with GBDT slightly outperforming others. In contrast, the BP (Backpropagation Neural Network) model shows the lowest R^2 and the highest errors (MAE, MSE, RMSE, and MAPE), suggesting inferior performance during training. Among the metrics, GBDT achieves the lowest MAE (3.136), MSE (17.785), and RMSE (4.217), demonstrating superior fitting quality. Other tree-based models, such as RF and ET, perform strongly but are less accurate than GBDT.

Table **13** evaluates the models on test data, revealing how well they generalize. CatBoost, XGBoost, and GBDT maintain strong R^2 values (above 0.87) and low error metrics, indicating good predictive capabilities. CatBoost achieves the lowest MAE (20.801) and high R2 (0.908), making it the top performer on test data. LightGBM and

XGBoost also show competitive performance, while BP struggles, with the highest MSE (3837.057) and RMSE (61.944), highlighting its poor generalization. The strong alignment between training and testing results for tree-based models suggests robust performance across data splits.

Table **14** highlights the importance of input features for each model. Across most models (Fig. **4**), Site EUI is the most significant feature, contributing over 82% of the predictive weight in all tree-based methods (and exceeding 93% for GBDT, RF, and ET). Floor Area and Year Built are moderately influential in LightGBM, while their impact is less pronounced in other models. Notably, facility type is of relatively low importance across all models. These results indicate that energy use intensity (Site EUI) is the dominant factor in model predictions. It is a direct reflection of the overall energy consumption efficiency of a building, particularly the operation of the heating, ventilation, and air conditioning (HVAC) system. The energy efficiency of the HVAC equipment, its maintenance status, and the insulation of the building envelope significantly affect the Site EUI, making it a key predictor of a building's primary energy use. In contrast, the effect of floor area on energy use is more muted, as reflected in the literature. For example, it has been shown that while a larger floor area generally implies higher total energy use, its effect may be masked by HVAC operating efficiency, usage patterns, and climatic conditions when normalized (i.e., how the EUI is calculated).

The main reason for the fit variability is due to the different algorithmic characteristics of the models and their ability to handle data features (Fig. **5**). Tree models (e.g., CatBoost, XGBoost, GBDT, etc.) capture nonlinear relationships by splitting the feature space and have a strong fitting ability to high-dimensional and unbalanced data, thus showing high accuracy in training and testing [39]. On the other hand, BP neural networks are sensitive to parameter initialization and optimization of the training process and are prone to fall into local optimums. At the same time, their processing of nonlinear features relies on larger sample sizes and tuning optimization, resulting in poorer fitting results [40]. In addition, there are also differences in the allocation of feature weights among the models. For example, the tree model pays more attention to Site EUI as a key feature. At the same time, BP cannot effectively capture the importance of the feature, which further widens the fitting differences among the models. Its limitations are reflected in its weak ability to process time-series data and difficulty capturing long-term dependencies. Alternative architectures, such as Long Short-Term Memory Networks (LSTM), can be considered to address this issue. By introducing memory units and gating mechanisms, LSTM can effectively handle the time-series characteristics of building energy consumption, such as seasonal variations, usage patterns, and the dynamic effects of weather factors. In addition, combining LSTM with CNN (Convolutional Neural Network) or Transformer architecture can further improve the model's prediction accuracy and enhance its ability to model complex nonlinear relationships.

4.3. Discussions

Physical modeling and data-driven approaches are two main approaches to building energy consumption prediction [41, 42]. Physical modeling relies on thermodynamic principles to simulate and analyze building energy

Table 12: Model evaluation train results.

Index	CatBoost	LightGBM	AdaBoost	XGBoost	GBDT	BP	RF	ET
R^2	0.997	0.931	0.952	0.996	0.999	0.656	0.984	0.985
MAE	5.610	17.173	24.038	6.447	3.136	56.237	10.008	9.518
MSE	53.156	1287.426	902.707	74.912	17.785	6443.479	306.552	272.184
RMSE	7.291	35.881	30.045	8.655	4.217	80.271	17.509	16.498
MAD	4.396	10.205	20.126	4.675	2.325	41.114	6.622	6.061
MAPE	0.159	0.383	0.691	0.177	0.090	1.700	0.213	0.205
EVS	0.997	0.931	0.952	0.996	0.999	0.657	0.984	0.986
MSLE	0.003	0.017	0.049	0.004	0.001	0.195	0.005	0.005

Table 13: Model evaluation test results.

Index	CatBoost	LightGBM	AdaBoost	XGBoost	GBDT	BP	RF	ET
R^2	0.908	0.871	0.845	0.891	0.876	0.618	0.875	0.889
MAE	20.801	21.907	26.747	20.663	22.031	43.821	23.409	21.742
MSE	922.194	1291.835	1552.812	1096.311	1248.673	3837.057	1256.819	1108.981
RMSE	30.368	35.942	39.406	33.111	35.337	61.944	35.452	33.301
MAD	14.721	14.602	19.170	13.284	13.932	30.573	16.847	15.498
MAPE	0.123	0.119	0.154	0.112	0.120	0.314	0.126	0.117
EVS	0.909	0.873	0.846	0.893	0.878	0.625	0.877	0.891
MSLE	0.024	0.022	0.036	0.021	0.025	0.122	0.024	0.021

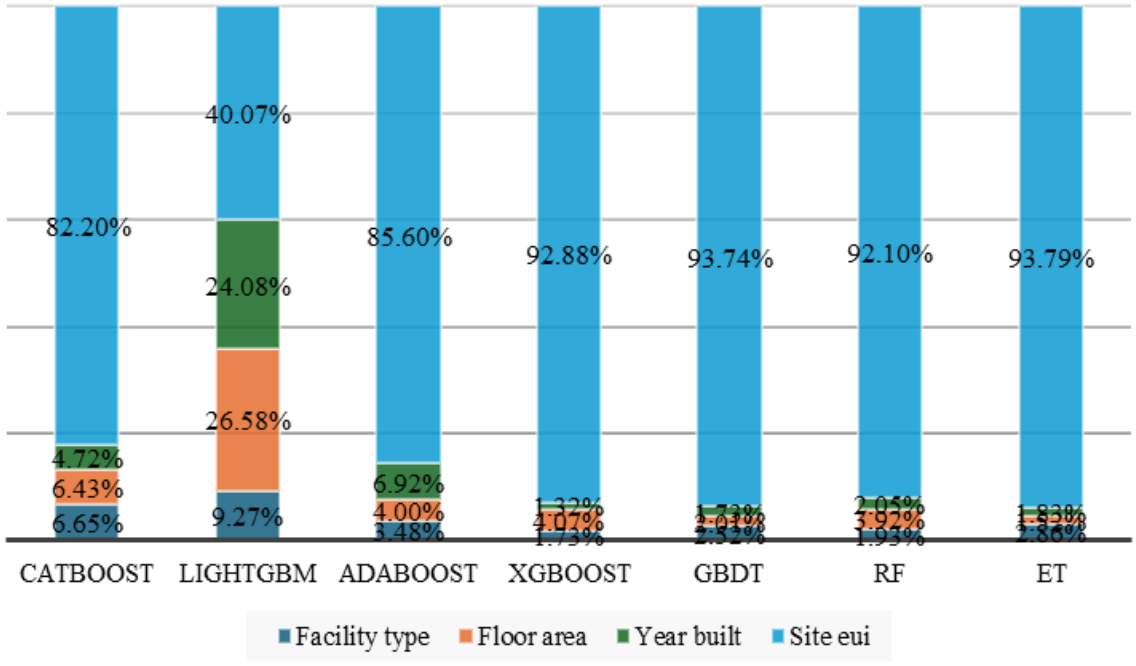


Figure 4: Comparison of the feature's weighting results.

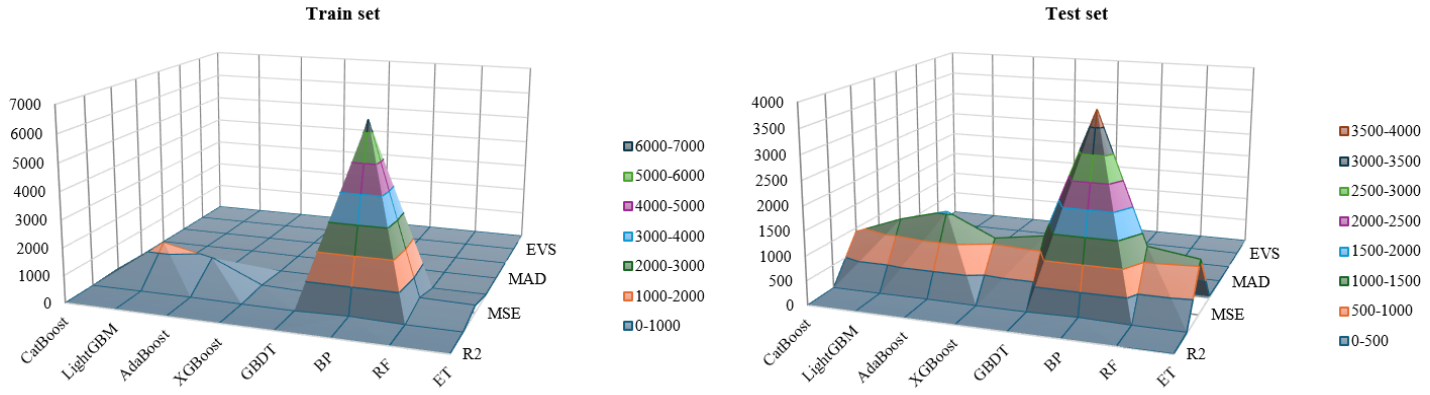


Figure 5: Comparison of the results of each model fit.

Table 14: Feature weighting results.

Index	CatBoost	LightGBM	AdaBoost	XGBoost	GBDT	BP	RF	ET
Facility type	6.65%	9.27%	3.48%	1.73%	2.52%	N/A	1.93%	2.86%
Floor area	6.43%	26.58%	4.00%	4.07%	2.01%	N/A	3.92%	1.52%
Year built	4.72%	24.08%	6.92%	1.32%	1.73%	N/A	2.05%	1.83%
Site eui	82.20%	40.07%	85.60%	92.88%	93.74%	N/A	92.10%	93.79%

consumption in detail. Commonly used building energy simulation software includes EnergyPlus®, eQuest®, and Ecotect® [43]. These software programs calculate the energy consumption of a building by inputting detailed building and environmental parameters such as building construction details, operation schedules, HVAC design information, climatic conditions, sky conditions, and solar/shading factors [44]. However, in real-world simulations, it is often difficult for users to obtain all the necessary details, which can affect the accuracy of the inputs [45, 46]. In contrast, data-driven approaches to building energy prediction do not require complex energy analysis or rely on detailed modeling of the building but instead, achieve energy prediction by learning from historical data [47, 48]. It is the latter approach that is used in this paper.

We recommend developing systems that visualize the energy consumption of each building in a city, allowing companies to quickly identify outliers (buildings that consume far more energy than expected even after adjusting for relevant predictors) [49, 50]. For example, they can target homes for potential retrofits or tiered pricing schemes [51]. For other end-users, an interface could be provided to enter their electricity and gas usage, as well as basic household information, to determine how their consumption compares to that predicted by the model for similar buildings [52]. Making users aware of their consumption in this way and relating it to consumption in similar buildings can produce behavioral changes that can lead to significant reductions in consumption.

This paper demonstrates that by learning from a large amount of historical energy consumption data, data-driven models (e.g., ML models) can capture the complex nonlinear relationships of building energy consumption, and their prediction accuracies outperform those of traditional physical models in many cases, especially when the data quality is high [53]. Second, compared to physical modeling that requires inputting many buildings' physical and environmental parameters, data-driven approaches are more rapid and easy to operate, making them more suitable for real-world engineering applications and rapid assessment scenarios [54]. Third, the data-driven approach can be flexibly adapted to predict energy consumption in different types of buildings, different climatic zones, and even different usage scenarios (e.g., commercial, residential, and industrial buildings) [55].

Site Energy Use Intensity (Site EUI) dominates EPEU because it directly reflects the overall energy consumption efficiency, especially the operation of heating, ventilation, and air conditioning (HVAC) [56, 57]. The energy efficiency of the HVAC equipment, its maintenance conditions, and the insulation of the building envelope can significantly affect Site EUI, which can thus become a predictive building primary energy use. In contrast, the effect of the floor area on energy consumption is relatively weak, as reflected in the literature [58, 59]. For example, it has been shown that while a larger floor area generally implies higher total energy use, its effect may be masked by HVAC operating efficiency, occupancy patterns, and climatic conditions when normalized (i.e., how the EUI is calculated) [60, 61].

5. Conclusion

The paper provides insights into the potential of machine learning algorithms for predicting building equivalent primary energy use (EPEU) using a comprehensive dataset for the Portland area. By systematically evaluating the performance of multiple ML models, the study finds that ensemble learning methods such as Random Forest (RF) and Gradient Boosting Machine (GBDT) show excellent results in energy prediction. At the same time, algorithms such as CatBoost and XGBoost also show high accuracy and robustness. These results provide valuable support for data-driven decision-making in building energy efficiency improvement and energy management. However,

the study also points out several directions that need further exploration. First, incorporating more dynamic factors (e.g., occupant behavior, real-time energy consumption data, and renewable energy integration) can help improve the model's predictive power. Second, expanding the scope of the study to a broader geographic area and diverse building types could enhance the applicability and generalizability of the findings. In the future, incorporating explainable artificial intelligence (XAI) techniques can help reveal key drivers of energy use, thus providing policymakers and building managers with a more straightforward basis for intervention. In conclusion, this study validates the broad application prospects of machine learning in building energy analysis. It provides a critical research foundation and direction for addressing urban sustainability challenges in the future.

It is worth noting that regional biases, such as Portland's mild climate, differ from those in other regions. Therefore, validation should be done in different climates, such as hot and humid or cold regions. Future research could integrate Internet of Things (IoT) data, such as real-time occupancy and indoor and outdoor climate parameters, to improve prediction accuracy. Some hybrid modeling approaches, such as combining physically driven models with machine learning (ML) methods can also be tried to better capture the dynamic characteristics of building energy consumption.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding

The authors declare that this research did not receive a specific grant from any public, commercial, or not-for-profit funding agency.

Acknowledgments

We thank all the anonymous reviewers for their valuable suggestions and the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy for data support.

Author Contributions

Yin Junjia collected the data, finished the data analysis, and completed the paper writing. Aidi Hizami Alias supervised the manuscript writing. Nuzul Azam Haron and Nabilah Abu Bakar encouraged and guided the work. All authors approved the last version.

References

[1] Gillingham KT, Huang P, Buehler C, Peccia J, Gentner DR. The climate and health benefits from intensive building energy efficiency improvements. Sci Adv. 2021; 7: 0947. https://doi.org/10.1126/sciadv.abg0947

[2] US Energy Information Administration (EIA). Total Energy Monthly Data - U.S. Energy Information Administration (EIA). Wwweiagov 2024. https://www.eia.gov/totalenergy/data/monthly/#consumption (accessed March 29, 2025).

[3] De Oliveira Matias JC, Devezas TC. Consumption dynamics of primary-energy sources: The century of alternative energies. Appl Energy 2007; 84: 763-70. https://doi.org/10.1016/j.apenergy.2007.01.007

[4] Paska J, Biczel P, Kłos M. Hybrid power systems – An effective way of utilising primary energy sources. Renew Energy. 2009; 34: 2414-21. https://doi.org/10.1016/j.renene.2009.02.018

[5] Papadis E, Tsatsaronis G. Challenges in the decarbonization of the energy sector. Energy. 2020; 205: 118025. https://doi.org/10.1016/j.energy.2020.118025

[6] Abdelrahman MM, Zhan S, Miller C, Chong A. Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. Energy Build. 2021; 242: 110885. https://doi.org/10.1016/j.enbuild.2021.110885

[7] Liu X, Gou Z. Occupant-centric HVAC and window control: A reinforcement learning model for enhancing indoor thermal comfort and energy efficiency. Build Environ. 2024; 250: 111197. https://doi.org/10.1016/j.buildenv.2024.111197

[8] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015; 349: 255-60. https://doi.org/10.1126/science.aaa8415

[9] Ai L, Muggleton SH, Hocquette C, Gromowski M, Schmid U. Beneficial and harmful explanatory machine learning. Machine Learn. 2021; 110: 695-721. https://doi.org/10.1007/s10994-020-05941-0

[10] Fathi S, Srinivasan R, Fenner A, Fathi S. Machine learning applications in urban building energy performance forecasting: A systematic review. Renew Sustain Energy Rev. 2020; 133: 110287. https://doi.org/10.1016/j.rser.2020.110287

[11] Amasyali K, El-Gohary N. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. Renew Sustain Energy Rev. 2021; 142: 110714. https://doi.org/10.1016/j.rser.2021.110714

[12] Olu-Ajayi R, Alaka H, Sulaimon I, Sunmola F, Ajayi S. Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. J Build Eng. 2022; 45: 103406. https://doi.org/10.1016/j.jobe.2021.103406

[13] Shapi MKM, Ramli NA, Awalin LJ. Energy consumption prediction by using machine learning for smart building: case study in Malaysia. Dev Built Environ. 2020; 5: 100037. https://doi.org/10.1016/j.dibe.2020.100037

[14] Pham A-D, Ngo N-T, Ha Truong TT, Huynh N-T, Truong N-S. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. J Cleaner Prod. 2020; 260: 121082. https://doi.org/10.1016/j.jclepro.2020.121082

[15] Wang R, Lu S, Feng W. A novel improved model for building energy consumption prediction based on model integration. Appl Energy. 2020; 262: 114561. https://doi.org/10.1016/j.apenergy.2020.114561

[16] Lei L, Chen W, Wu B, Chen C, Liu W. A building energy consumption prediction model based on rough set theory and deep learning algorithms. Energy Build. 2021; 240: 110886. https://doi.org/10.1016/j.enbuild.2021.110886

[17] Letham B, Karrer B, Ottoni G, Bakshy E. Constrained bayesian optimization with noisy experiments. Bayesian Anal. 2019; 14: 495-519. https://doi.org/10.1214/18-ba1110

[18] Zulfiqar M, Gamage KAA, Kamran M, Rasheed MB. Hyperparameter optimization of bayesian neural network using bayesian optimization and intelligent feature engineering for load forecasting. Sensors. 2022; 22: 4446. https://doi.org/10.3390/s22124446

[19] Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition 2015; 48: 2839-46. https://doi.org/10.1016/j.patcog.2015.03.009

[20] Satria A, Sitompul OS, Mawengkang H. 5-fold cross validation on supporting k-nearest neighbour accuration of making consimilar symptoms disease classification. IEEE Xplore. 2021; 1: 1-5. https://doi.org/10.1109/IC2SE52832.2021.9792094

[21] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020; 7: 94. https://doi.org/10.1186/s40537-020-00369-8

[22] Zhou F, Pan H, Gao Z, Huang X, Qian G, Zhu Y, *et al*. Fire prediction based on catboost algorithm. Math Problems Eng. 2021; 2021: 1-9. https://doi.org/10.1155/2021/1929137

[23] Ju Y, Sun G, Chen Q, Zhang M, Zhu H, Rehman MU. A model combining convolutional neural network and lightGBM algorithm for ultra-short-term wind power forecasting. IEEE Access. 2019; 7: 28309-18. https://doi.org/10.1109/access.2019.2901920

[24] Onoja M, Jegede A, Blamah N, Olawale AV, Omotehinwa TO. EEMDS: efficient and effective malware detection system with hybrid model based on xceptionCNN and lightGBM algorithm. J Comput Soc Inform. 2022; 1: 42-57. https://doi.org/10.33736/jcsi.4739.2022

[25] Cao Y, Miao QG, Liu JC, Gao L. Advance and prospects of adaboost algorithm. Acta Automatica Sinica. 2014; 39: 745-58. https://doi.org/10.3724/sp.j.1004.2013.00745

[26] Hastie T, Rosset S, Zhu J, Zou H. Multi-class AdaBoost. Stat Interface. 2009; 2: 349-60. https://doi.org/10.4310/sii.2009.v2.n3.a8

[27] Hah DW, Kim YM, Ahn JJ. A study on KOSPI 200 direction forecasting using XGBoost model. J Korean Data Inform Sci Soc. 2019;30: 655-69. https://doi.org/10.7465/jkdi.2019.30.3.655

[28] Oh J-Y, Ham D-H, Lee Y-G, Kim G. Short-term load forecasting using XGBoost and the analysis of hyperparameters. Transac Korean Inst Electr Eng. 2019; 68: 1073-8. https://doi.org/10.5370/kiee.2019.68.9.1073

[29] Zhang Z, Jung C. GBDT-MO: gradient-boosted decision trees for multiple outputs. IEEE Trans Neural Netw Learn Syst. 2021; 32: 3156-67. https://doi.org/10.1109/tnnls.2020.3009776

[30] Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci (Basel). 2018; 8: 689. https://doi.org/10.3390/app8050689

[31] Jin T, Zhou ZY. Leakage detection method for piping network based on BP Neural Network. Appl Mech Mater. 2013; 470: 738-42. https://doi.org/10.4028/www.scientific.net/amm.470.738

[32] Jin J. Fault diagnosis of coal equipment based on dynamic fuzzy neural network and BP Neural Network. Int J Hybrid Inf Technol. 2016; 9: 275-82. https://doi.org/10.14257/ijhit.2016.9.7.25

[33] Afanador NL, Smolinska A, Tran TN, Blanchet L. Unsupervised random forest: a tutorial with case studies. J Chemometr. 2016; 30: 232-41. https://doi.org/10.1002/cem.2790

[34] Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl. 2024; 237: 121549. https://doi.org/10.1016/j.eswa.2023.121549

[35] Baldini G. Mitigation of adversarial attacks in 5g networks with a robust intrusion detection system based on extremely randomized trees and infinite feature selection. Electronics. 2024; 13: 2405. https://doi.org/10.3390/electronics13122405

[36] Li X, Jiang S, Wang X, Wang T, Zhang S, Guo J, *et al*. XCO2 super-resolution reconstruction based on spatial extreme random trees. Atmosphere. 2024; 15: 440. https://doi.org/10.3390/atmos15040440

[37] Mao W. Composition analysis and identification of ancient glass objects using regression and clustering algorithms. Highl Sci Eng Technol. 2023; 35: 6-11. https://doi.org/10.54097/hset.v35i.7016

[38] Li M, Zhou Q, Han X, Lv P. Prediction of reference crop evapotranspiration based on improved convolutional neural network (CNN) and long short-term memory network (LSTM) models in Northeast China. J Hydrol. 2024; 645: 132223. https://doi.org/10.1016/j.jhydrol.2024.132223

[39] Zhou J, Su Z, Hosseini S, Tian Q, Lu Y, Luo H, *et al*. Decision tree models for the estimation of geo-polymer concrete compressive strength. Math Biosci Eng. 2023; 21: 1413-44. https://doi.org/10.3934/mbe.2024061

[40] Zhang L, Wang F, Sun T, Xu B. A constrained optimization method based on BP neural network. Neural Comput Appl. 2016; 29: 413-21. https://doi.org/10.1007/s00521-016-2455-9

[41] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev. 2018; 81: 1192-205. https://doi.org/10.1016/j.rser.2017.04.095

[42] Heidarinejad M, Guillermo J, Wentz JR, Rekstad NM, Spengler JD, Jelena Srebric. Actual building energy use patterns and their implications for predictive modeling. Energy Convers Manag. 2017; 144: 164-80. https://doi.org/10.1016/j.enconman.2017.04.003

[43] Zhao H, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev. 2012; 16: 3586-92. https://doi.org/10.1016/j.rser.2012.02.049

[44] Qiao Q, Yunusa-Kaltungo A, Edwards RE. Towards developing a systematic knowledge trend for building energy consumption prediction. J Build Eng. 2020; 35: 101967. https://doi.org/10.1016/j.jobe.2020.101967

[45] Kontokosta CE, Tull C. A data-driven predictive model of city-scale energy use in buildings. Appl Energy. 2017; 197: 303-17. https://doi.org/10.1016/j.apenergy.2017.04.005

[46] Wang Z, Srinivasan RS. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renew Sustain Energy Rev. 2017; 75: 796-808. https://doi.org/10.1016/j.rser.2016.10.079

[47] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy. 2014; 127: 1-10. https://doi.org/10.1016/j.apenergy.2014.04.016

[48] Zhang Y, O'Neill Z, Dong B, Augenbroe G. Comparisons of inverse modeling approaches for predicting building energy performance. Build Environ. 2015; 86: 177-90. https://doi.org/10.1016/j.buildenv.2014.12.023

[49] Kolter J, Ferreira J. A large-scale study on predicting and contextualizing building energy usage. Proc AAAI Conf Artif Intell. 2011; 25: 1349-56. https://doi.org/10.1609/aaai.v25i1.7806

[50] Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl Energy. 2015; 160: 153-63. https://doi.org/10.1016/j.apenergy.2015.08.126

[51] Xiaoxiang Q, Junjia Y, Haron NA, Alias AH, Teik Hua L, Abu Bakar N. Status, challenges and future directions in the evaluation of net-zero energy building retrofits: a bibliometrics-based systematic review. Energies. 2024; 17: 3826. https://doi.org/10.3390/en17153826

[52] Xiaoxiang Q, Junjia Y, Haron NA, Alias AH, Law TH, Nabilah AB. Customer perceived value theory and PSO-LightGBM algorithm-based approach to evaluating satisfaction factors with Net-zero energy building retrofits. Edelweiss Appl Sci Technol. 2025; 9: 2508-30. https://doi.org/10.55214/25768484.v9i3.5835

[53] Junjia Y, Alias AH, Haron NA, Bakar NA. Intelligent construction risk management through transfer learning: trends, challenges, and future strategies. Artif Intell Evol. 2024; 6: 1-16. https://doi.org/10.37256/aie.6120255255

[54] Junjia Y, Alias AH, Haron NA, Bakar NA. Machine learning algorithms for safer construction sites: Critical review. Build Eng. 2024; 2: 544. https://doi.org/10.59400/be.v2i1.544

[55] Junjia Y, Alias AH, Haron NA, Abu Bakar N. Deep learning for safety risk management in modular construction: Status, strengths, challenges, and future directions. Autom Constr. 2025; 169: 105894. https://doi.org/10.1016/j.autcon.2024.105894

[56] Karatasou S, Santamouris M, Geros V. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy Build. 2006; 38: 949-58. https://doi.org/10.1016/j.enbuild.2005.11.005

[57] Wang Z, Wang Y, Zeng R, Srinivasan RS, Ahrentzen S. Random Forest based hourly building energy prediction. Energy Build 2018; 171: 11-25. https://doi.org/10.1016/j.enbuild.2018.04.008

[58] Wang Z, Wang Y, Srinivasan RS. A novel ensemble learning approach to support building energy use prediction. Energy Build. 2018; 159: 109-22. https://doi.org/10.1016/j.enbuild.2017.10.085

[59] Korolija I, Marjanovic-Halburd L, Zhang Y, Hanby VI. UK office buildings archetypal model as methodological approach in development of regression models for predicting building energy consumption from heating and cooling demands. Energy Build. 2013; 60: 152-62. https://doi.org/10.1016/j.enbuild.2012.12.032

[60] Deng H, Fannon D, Eckelman MJ. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build. 2018; 163: 34-43. https://doi.org/10.1016/j.enbuild.2017.12.031

[61] Golafshani E, Chiniforush AA, Zandifaez P, Ngo T. An artificial intelligence framework for predicting operational energy consumption in office buildings. Energy Build. 2024; 317: 114409. https://doi.org/10.1016/j.enbuild.2024.114409