Energy 262 (2023) 125468

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence

Yan Zhang [a], Bak Koon Teoh [a], Maozhi Wu [b], Jiayu Chen [c], Limao Zhang [d,*]

[a] School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[b] Hubei Jianke Technology Group, 430223, Wuhan, China
[c] Department of Architecture and Civil Engineering, City University of Hong Kong, B6322, Tat Chee Avenue, Kowloon, Hong Kong
[d] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Hongshan District, Wuhan, Hubei 430074, China

ARTICLE INFO

Keywords:
Urban morpho-blocks
Data-driven estimation
Light gradient boosting machine
Explainable AI
Building energy consumption

ABSTRACT

Energy consumption prediction is an integral part of planning and controlling energy used in the building sector which accounts for 40% of the global energy consumption and a significant portion of greenhouse gas emissions. However, very few studies focused on the combined effect of building characteristics, building geometry, and urban morphology on energy performance. Such a research gap is addressed in this study by developing an explainable deep learning model. Our model uses Light Gradient Boosting Machine integrated with the SHapley Additive exPlanation algorithm, so as to provide insights into the feasibility of using machine learning-based models for energy performance prediction of buildings. With the proposed eXplainable Artificial Intelligence model, this study successfully predicts energy usage and greenhouse gas emissions of residential buildings, as well as identifies the most influential variables and evaluates their relative importance. A case study based on Seattle's data is used to verify the proposed framework, and some conclusions can be drawn: (1) Urban morphology and building geometry have significant effects on evaluating the building energy consumption and greenhouse gas emissions, as the accuracy of predicted result improve 33.46% compared with only considering building characteristics; (2) The total gross floor area and natural gas are identified as the most influential factors for energy consumption and GHG emissions, respectively; (3) The proposed model is examined to be an accurate method with the R^2 of 0.8435 on average, comparing with the other approaches, such as the eXtreme Gradient Boosting, Random Forest, and Support Vector Regression. The main contributions of this research lie in that (a) a comprehensive structure integrated with building characteristics, building geometry, and urban morphology is established to forecast the energy use and greenhouse gas emissions; (b) an explainable artificial intelligence model incorporated with the SHapley Additive exPlanation algorithm into Light Gradient Boosting Machine has been proved to achieve an accurate prediction of the energy performance of residential buildings.

1. Introduction

The buildings and buildings construction sectors account for one-third of total global final energy consumption and nearly 15% of CO_2 emissions. With the rapid growth of the population and the increasing need for energy-consuming appliances such as air conditioning in emerging countries, the energy demand for buildings is projected to continue to climb. Many cities throughout the world have set targets to enhance energy efficiency and cut greenhouse gas (GHG) emissions, aiming to mitigate the environmental impact and achieve long-term sustainable development. Therefore, accurate quantifying of building energy consumption has become a prominent concern. To better predict energy consumption and achieve better energy efficiency, experts from academics and industry have drawn their attention to reveal the opportunities for energy-saving measures, particularly in individual buildings [1]. Existing buildings energy-related studies are often focused on economics [2], climate change [3], users behaviors [4], and construction and layout [5], and sub-level components such as lighting, HVAC (Heating, Ventilating, and Air-Conditioning) systems [6]. However, limited studies have integrated morphology and building layout to assess and predict energy usage and GHG emissions.

The dynamic and complex nature of building systems has made the prediction of building energy consumption and GHG emissions more difficult. Especially for individual buildings, where the energy

* Corresponding author.
E-mail addresses: yan007@e.ntu.edu.sg (Y. Zhang), bakkoon.teoh@ntu.edu.sg (B.K. Teoh), jiaychen@cityu.edu.hk (J. Chen), zlm@hust.edu.cn (L. Zhang).

https://doi.org/10.1016/j.energy.2022.125468
Received 15 May 2022; Received in revised form 23 August 2022; Accepted 14 September 2022
Available online 19 September 2022

consumption behavior of different buildings is difficult to collect and time-consuming due to privacy, surrounding environment, and various physical characteristics [7]. Firstly, building characteristics and physical features, including building layout, building shape, building age, etc., are considered as the potential influence factors of energy consumption. For example, Wang and Greenberg [8] stated that the window position can improve the energy efficiency and thermal comfort of the building even in a hot-humid climate. Sonta et al. [9] found that the design of the building layout could reduce the lighting energy consumption by 5%–6% of the total energy usage for office space. Heydari et al. [10] confirmed that the design parameters, such as the facade orientation, window-to-wall ratio, and building envelope, had a significant effect on the energy performance and comfort of buildings. Hemsath and Bandhosseini [11] conducted a study to examine the building geometry's impact on energy use. They analyzed the width-to-length ratio and surface-to-volume ratio, and the results showed that the geometric form factor in some cases had a more significant impact on buildings' energy performance than the material used in the building envelope. Secondly, urban buildings are more susceptible to their surrounding objects, since the external environment may affect the buildings' energy demand due to building density-induced micro-climate changes. Park and Cho [12] investigated the relationship between green space and building density vs. height, where the cooling distance of the green field can be found within 120 m with a cooling effect of 3.0 K. Srebric et al. [13] emphasized that the neighborhood and streets could affect the distribution of the heat transfer process among the buildings. They recommended that the tall buildings along with the broader streets may relieve the energy burden more than shorter buildings along with narrower streets.

Urban morphology and building geometry are two of the vital influence factors for urban buildings' energy assessment, since they significantly affect the outdoor thermal performance, building environment, and energy consumption [14]. Many studies have attempted to investigate the influence factors of urban morphology and building geometry, including building density, floor area ratio, etc. According to Strømann-Andersen and Sattrup [15]; building density had an influence on natural lighting and passive solar gains, and it also increased energy consumption for lighting, cooling, and heating by up to 30% for workplaces and 19% for dwellings. on the natural lighting and passive solar gains of buildings and further impacted energy usage for lighting, cooling, and heating by up to 30% in offices and 19% in dwellings. Güneralp et al. [16] conducted bottom-up research and the results showed that the trade-offs of both urban density and building energy efficiency would reduce building energy consumption and improve climate change. Shashua-Bar et al. [17] identified that tree coverage, street deepening, and street ventilation played important roles in moderating the outdoor temperature. To evaluate the energy consumption, Furthermore, previous studies have attempted to optimize the urban morphology based on building shape, aiming to achieve a more efficient way of energy use. Alznafer [18] found that the height/width (H/W) ratio provided more outdoor shading, resulting in energy savings at the indoor level. In addition, NE-SW canyon orientation showed less energy consumption in hot climate areas such as the 'Arabian Gulf' region.

Morphological information could be generated using Geographic Information Systems (GIS). For instance, Ma and Cheng [19] deployed GIS to determine the distance between buildings and the shore in terms of natural cooling, the density value of population and traffic, and the Normalized Difference Vegetation Index. Tong et al. [20] investigated how landscapes and building geometry elements affect air temperature in Tianjin, China. In essence, the building energy consumption assessment requires efforts to integrate the urban morphology and physical characteristics. However, limited studies tried to incorporate urban morphology and building geometry to explore the relationship between urban morphology and building energy consumption.

Given this background, this study attempts to evaluate the building energy performance by considering the urban morphology, building geometry, and building features. Unlike the previous black-box type of Machine learning methods that merely present the details, we applied an eXplainable artificial intelligence (XAI) model via using Light Gradient Boosting Machine (LightGBM) integrated with the SHapley Additive explanation (SHAP) method to quantitively predict how different characteristics influence the building energy consumption. Moreover, we investigate the importance of factors with regard to urban morphology, surrounding environment, and building physical features, which may provide valuable information for the other regions and countries where they may face similar concerns and issues. The integrated SHAP approach enhances the interpretability of the LightGBM model by determining the influence of the factors for accurate prediction of the target and investigating the interactions and dependencies amongst the influential variables with the objectives. The findings should shed light on these questions: (1) what is the relationship between urban morphology and individual building performance; (2) which factors have a significant impact on the building energy performance.

The rest of this research is organized as follows. Related data-driven studies are reviewed in Section 2, which is followed by the methodology and model development section. Section 4 illustrates the application of the proposed framework. Section 5 demonstrates the case study results, while Section 6 presents some inferences and discusses future works.

2. Related studies

Since many features influence the prediction of building energy consumption, the relationships between each component and the objective will need to be considered using regression models based on training data obtained from observations [21]. There are various machine learning methods used in the assessment of energy levels, including the prediction of building energy consumption in different areas. Many previous pieces of research have predicted the building energy performance using machine learning methods, which plays a crucial role as it helps to predict the future values of objectives by utilizing the observed values in the past [22]. In summary, there are three main supervised learning algorithms in machine-learning techniques for evaluating building energy performance, these include Artificial Neural Network (ANN), Support Vector Machine (SVM), and Gaussian distribution regression models [23]. Specifically, Alam et al. [24] developed the ANN model to predict the heating and cooling loads considering the surface area, wall area, roof area, overall height, and orientation by using the 768 residential buildings dataset. Li et al. [25] proposed an ANN-based fast buildings energy consumption prediction method, to be used during the building design stage, which shifts the complexity of energy performance issues into several energy consumption problems with multiple simplified blocks. Ascione et al. [26] integrated the ANN method into EnergyPlus simulation to analyze the building energy efficiency retrofit, where the result shows a very satisfactory reliability and a substantial reduction of the required computational burden (around 98%).

As SVMs display high robustness and have the ability to solve nonlinear problems and require less number of data samples, where the number of hyper-parameters is less than ANN [23]. Massana et al. [27] compared SVM, ANN, and multiple linear regression (MLR) in the short-term prediction of non-domestic buildings' electricity demand and concluded that SVM provided higher accuracy and lower computational cost. Li et al. [28] used SVM to predict electricity consumption by considering 15 building envelope features from 59 different cases in the long term. Shao et al. [29] applied SVM to find the relevant laws of the operating data and total energy consumption of the air conditioning unit for hotel buildings. Liu et al. [30] examined the influence of energy consumption, climatic change, and time-cycle factors on public buildings via the SVM method. Despite the extraordinary efforts of some of the above research efforts in data-driven building energy prediction, the complexity of building energy consumption results in the continued

challenges in the accuracy and efficiency of the assessment process. Hence, the Gaussian process (GP) regression method is deemed as an alternative approach to analyzing the uncertainties in building energy. Zeng et al. [31] discussed building electricity energy use and assisted in the utilization of energy storage devices to reduce the peak energy demands and energy supply risks. Yoon and Moon [32] identified the influence factors and provided a more accurate estimation of energy consumption via adopting GP regression. Yuan et al. [33] introduced the Gaussian distribution function to generate new samples around a similar sample and combined a novel sample data selection method (SDSM) to predict heating energy consumption. However, most of the machine learning methods mentioned above are regarded as "black-box" methods due to limited interpretability. These non-interpretable machine learning methods analyzed the data via internal logic in the black box that is hidden and cannot be understood by users, raising the risk of inducing wrong decisions.

To avoid the risk of non-interpretable AI models, many researchers adopt explainable AI in the energy domains, such as the Decision Tree (DT) model, Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). These tree-based models have received much attention in academic circles, as they explain the black-box models to some degree and are more flexible [34]. For example, Tso and Yau [35] applied DT to identify the information on the importance of significant driving forces, including household characteristics and appliance ownership, for buildings' energy consumption. However, the primary concern of DT is the overfitting issues that make its prediction not robust enough. Hence, Random Forest (RF) is proposed by Breiman [36]; which has been popularly utilized to tackle regression and classification problems. Ahmad et al. [37] suggested that the RF can be used as a variable selection tool for HVAC electricity consumption prediction. Meanwhile, the gradient boosted tree (GBDT) model integrates many tree models, which are widely used in different building energy consumption prediction cases to assess the energy demand level of the targeted study area and further identify the energy consumption pattern. Gradient boosting is a powerful and effective machine learning algorithm capable of developing models with excellent predictability. Among numerous machine learning techniques implemented in data prediction, Light Gradient Boosting Machine (LightGBM) is an effective and useful approach due to its ability to handle nonlinear problems in practice [38]. LightGBM has been proven to have up to 20 times faster training speed on the same amount of data set compared to the implementation of other GB algorithms such as XGBoost. With the advantages of LightGBM, this research, thus, employed this state-of-the-art approach to analyze and process data in a much shorter time.

3. Methodology

To investigate the urban buildings and predict the energy consumption and GHG emissions, a LightGBM method is applied in this research. Fig. 1 demonstrates the workflow of the prediction process. The latest individual building dataset, including the location and building types, is firstly identified, and the urban morphology is extracted based on the city blocks and building geometry. Then, the influential variables are identified and divided into three different categories for analyzing building energy consumption and GHG emissions. An efficiency LightGBM approach is employed to examine the effect of the urban morphology, where the accuracy of the result is validated by statistical indices. Meanwhile, the importance of urban and building characteristics in determining energy consumption and GHG emissions is evaluated, which may contribute to policy implementation on the users' and planners' sides.

3.1. Urban morphological analysis

Urban morphological parameters are highly associated with building geometry and distribution, which are measured by locations, height, and the type of buildings. It can affect the energy consumption and GHG emissions in a specific region with the consideration of the surrounding environment and the population density. Combining these different variables can reflect the unique regional building, whose characteristics can be calculated and normalized by standard procedures that make the cross-regional analysis possible. Advanced technology development in Geographic Information Systems (GIS) and Building Information Modeling (BIM) have enabled the construction of aggregated datasets

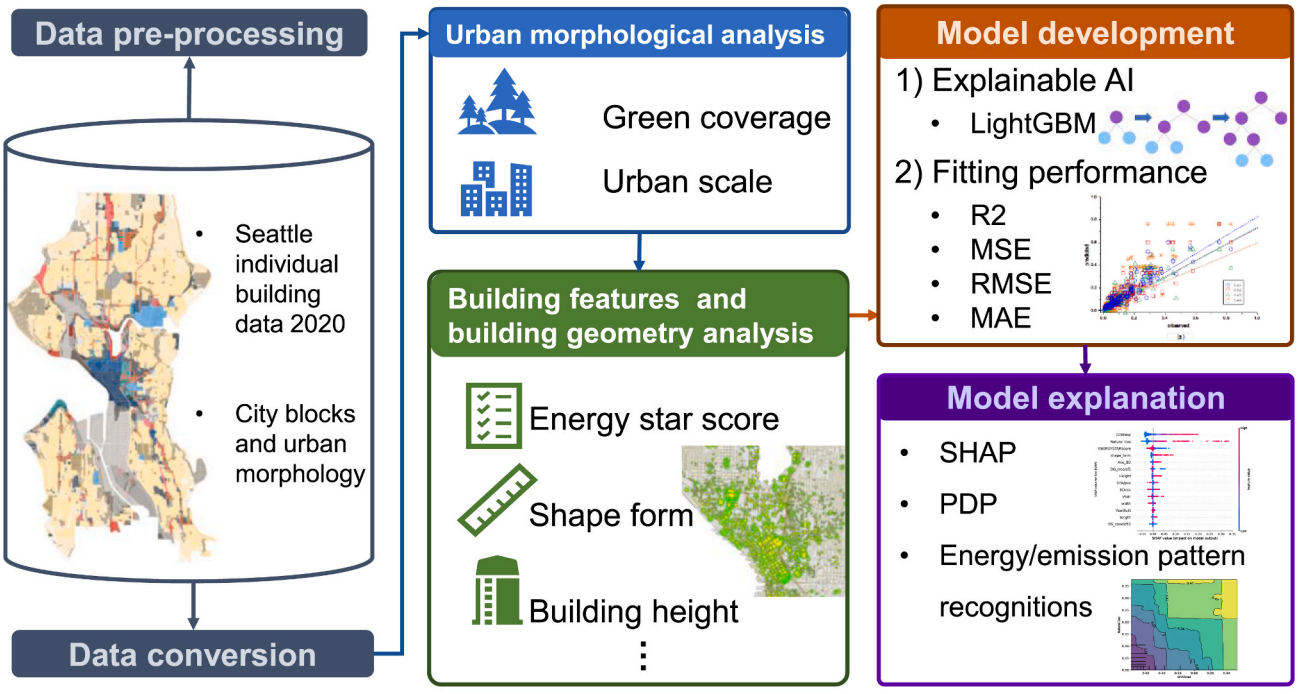

Fig. 1. Framework of the proposed method for estimation of building energy consumption and GHG emissions.

and the quantification of key physical building attributes of modern cities. By extracting building information from these models, this study uses building morphology as a parameter to evaluate and optimize building energy consumption and GHG emissions, taking into account the building's characteristics and the influence of its surroundings.

The potential influence factors of individual buildings regarding the urban morphology contain building density, road networks, tree canopy coverage, etc. To be specific, the density of the street is one of the significant features of urban scale. The street layout and building disposition design affect the building block prevailing wind, where the street canyon effect may change the temperature, wind speed, and wind direction to influence the buildings' energy demand [39]. Wei et al. [40] investigated the potential of saving buildings' energy with different vegetation types, and the results indicated that the effect of energy conservation on cooling load by shading from trees was more significant than that by evapotranspiration from lawns. Therefore, green coverage is determined as one of the critical factors affecting building energy use. Fahmy et al. [41] simulated the micro-climate for both indoor thermal comfort and energy consumption with the consideration of urban canopy green coverage. Results showed that the green coverage could improve the outdoor micro-climate and mitigate the residential building energy consumption. Similarly, Ouyang et al. [42] found that green cooling efficiency increases at the lower green coverage and decreases with green abundance increasing until a maximum of 20–30%. Besides, building density and building coverage ratio (BCR) play vital roles in urban building energy consumption, since they affect the shadow between buildings and ventilation efficiency in urban canyons [43,44]. The building coverage ratio is defined as the ratio of the building coverage area (i.e. the area of building footprint) to the size of the land lot in Eq. (1) and Table 1.

$$BCR = \frac{S_{bld}}{S_{Land}} \tag{1}$$

To initialize the urban morphology analysis, the building geometry and geospatial information need to be pre-processed so that they can be extracted based on the morphological features. Firstly, the individual building dataset, including the building age, energy stat score, and GFA, is obtained from the government open data websites (https://data.seattle.gov/). Then, the information on city blocks and road networks is collected from OpenStreetMap.org. This study adopted Geopandas, a python package, to read the original GeoJSON building dataset from OpenStreetMap (OSM) and the government website. The imported datasets were stored as a Geopandas object, which follows the GeoDataframe for further analysis. There are two main steps for extracting and processing data: (1) downloading the target area and extracting the fundamental geometry and footprint data from the open dataset and constructing the building footprint; (2) screening and aligning building footprints to incorporate geographic features and generate the new layer for future analysis by GIS. Fig. 2 shows an example of the city dataset for building layout extraction. The locations of the building are represented by dots in Fig. 2 (a), where the zoomed-in area is the city center. Fig. 2 (b) is the satellite image of the building layouts in the city center from Google Earth, and the extracted 2D building layouts colored by the yellow polygon can match the real location perfectly as shown in Fig. 2 (c).

Table 1
Urban morphological variables and corresponding descriptions.

Variables	Description	Reference
BCR	Building coverage ratio	Ouyang et al. [42]
S_{bld}	Building coverage area/footprint	Ouyang et al. [42]
S_{land}	The area of the building land lot	Ouyang et al. [42]

3.2. Building features and geometry analysis

Buildings' physical features and their geometry should be considered to evaluate the individual buildings' energy performance. For the building's physical features, the total gross floor area (GFA), natural gas, energy star score, and year of build are used to predict the energy usage and GHG emissions in this study. It is common that the large total GFA in the block consumes more electric energy at a building level so as to increase the volume of energy use [45]. Deb and Lee [46] pointed out that the age of buildings affected building energy performance, since the age of the HVAC systems is inefficient and needs retrofitting. Another factor, namely the energy star score, is determined as the fair benchmarking on the energy performance introduced by the USA, which processes a statistical analysis between peer buildings and normalizes their features. Dahlan et al. [47] first applied the energy star score approach in Asian areas to estimate hospital buildings. Natural gas in buildings belongs to a genre of energy resources that supports space heating, domestic hot water, and cooking. For example, Wei et al. [48] collected natural gas as one of the databases of building energy consumption and carbon dioxide emission in residential buildings.

From a building geometrical perspective, the influential factor, including building height, shape form factor, vertical to horizontal ratio, length of the building, and width of the building, of the building energy consumption and GHG emissions will be evaluated in this study by Eq. (2)~(4). For instance, Shareef [49] compared different scenarios with building height diversity on energy consumption, aiming to find the effect of variation in the building height on both outdoor microclimate and indoor energy consumption. Ma et al. [50] defined the concept of shape form by utilizing the minimum circumscribed circle to represent a morpho-block. Hemsath and Bandhosseini [11] conducted a sensitivity analysis of the building's geometry design decision with the consideration of vertical and horizontal proportional relationships. Aldawoud [51] analyzed that atrium-shaped buildings with high length-to-width ratios are less energy efficient than other geometries. However, the majority of existing studies lack efforts to incorporate urban morphology regarding the energy performance assessment. Thus, this research proposes a comprehensive structure of building energy prediction by integrating the urban morphology, building physical, and geometrical features. Fig. 3 illustrates the framework of the proposed prediction structure and Fig. 4 shows some details of selected factors.

$$R_{VH} = \frac{\sum_{i=1}^{N} BF_i}{A_{block}} \tag{2}$$

$$R_{BC} = \frac{\sum_{i=1}^{N} BA_i}{A_{block}} \tag{3}$$

$$F_{block} = \frac{A_{block}}{A_{circle}} \tag{4}$$

where Table 2 shows the descriptions of the parameters, including R_{VH} represents buildings vertical to horizontal ratio, BF_i is the facade area of building i, A_{block} stands for the block area, R_{BC} denotes the building coverage ratio, BA_i means the footprint area of building I, F_{block} is the block shape factor, and A_{circle} is the area of the minimum circumscribed circle of a specific building, which is expressed as $A_{circle} = \pi \bullet R^2_{circle}$.

3.3. LightGBM model development

Light Gradient Boosting Machine (LightGBM) is a popular model released in 2016 that uses a boosting algorithm to combine several week learners into a more accurate model [54]. Unlike the traditional generation approach of trees in Gradient Boosting Machine (GBM) that uses a leaf-wise generation strategy with parallel learning, as shown in Fig. 5

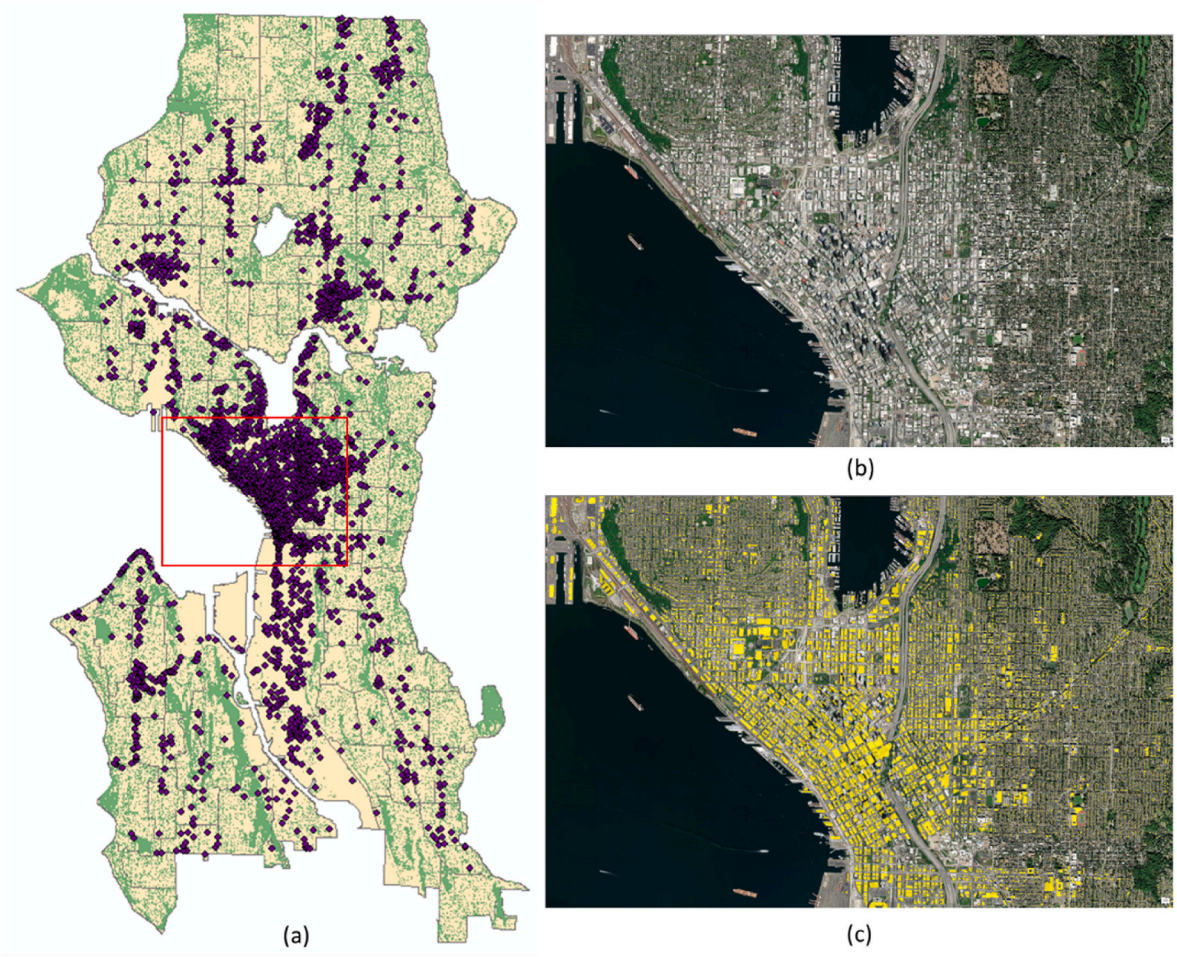

Fig. 2. Example of the result of building layout extraction: (a) location and layout of Seattle building; (b) building layout in satellite map; and (c) matching from the extracted building layout (yellow shadow) to satellite building map.

(a), resulting in an overfitting result, the LightGBM applies a leaf-wise generation that reduces the training error and improves the accuracy by using a histogram-based algorithm as shown in Fig. 5 (b) [55]. The strategy of lightGBM using leaf-wise-tree growth based on the node can majorly reduce the error, since the leaf-wise-tree algorithm has fewer tree nodes than the level-wise-tree algorithm with the same tree depth. Therefore, the training process could be significantly accelerated when dealing with a large dataset. Besides, lightGBM also provides other algorithms, including histogram-based, gradient-based one-side sampling (GOSS), and exclusive feature bundling (EFB) algorithms, to speed up the training process. Specifically, the histogram-based algorithm converts the sorted data set of parameters in the input layer into a histogram with a specified number of data intervals or bins, which allows the LightGBM model to require lower memory consumption while significantly speeding up the training speed. GOSS excludes data instances with small gradients and uses the remainder to estimate information. Since data with large gradients are more critical, information can be evaluated quickly and accurately, even from small-scale datasets. Furthermore, EFB bundles mutually exclusive variables and processes them to reduce the number of variables. The number of variables can be effectively reduced by bundling and processing variables that rarely have simultaneous nonzero values without significantly impairing accuracy. Thus, LightGBM achieves good performance with short training times for handling large datasets no matter whether they are sparse or not.

In this study, the LightGBM algorithm approach is proposed to predict the building energy consumption of Seattle. The main influence factors, where strong nonlinear interactions exist among them, are the components of the input layer of the proposed model. Apart from the input variables, a series of hyperparameters should be defined to ensure the model's performance. A GridSearchCV function is applied in this study to identify the optimal combination of the hyperparameters, which tests all possible combinations and compares a cross-validation process, finding the best hyperparameter combinations. The LightGBM model was referred from the open-source of Microsoft and the codes are run on the Python 3.8 environment.

After obtaining the prediction results from the constructed LightGBM regression model, the results are compared with the original data records to further evaluate the regression performance of the developed model in building energy consumption. The accuracy of the LightGBM model is examined based on four statistical indices, which are the coefficient of determination (R^2), the mean square error (MSE), the root mean square error (RMSE), and the mean absolute error (MAE) to analyze the fitting quality of predictions to the original observations. The coefficient of determination given in Eq. (5), and the value of R^2 with a range from 0 to 1 indicates the goodness of fit of a model, where a higher R^2 represents a good model performance in prediction. If the R^2 value is above 0.7, this value is generally considered strong effects [56, 57]. The rest of the three indices MSE, RMSE, and MAE calculated by Eqs. (6)–(8) are used to compute the difference between predicted and observed values. MSE measures the errors for gradient computation, while RMSE further calculates the standard deviation of the errors, whereas MAE represents the average magnitude of the errors in the predicted results. In general, smaller MSE, RMSE, and MAE are preferred

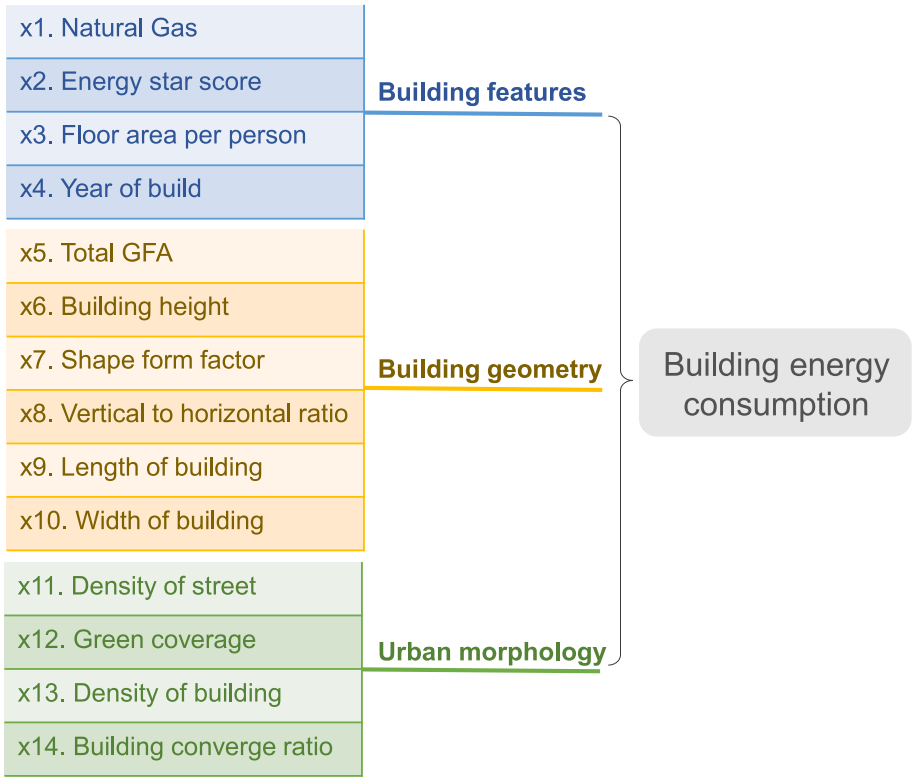

Fig. 3. Structure of the influential factors for building energy consumption prediction.

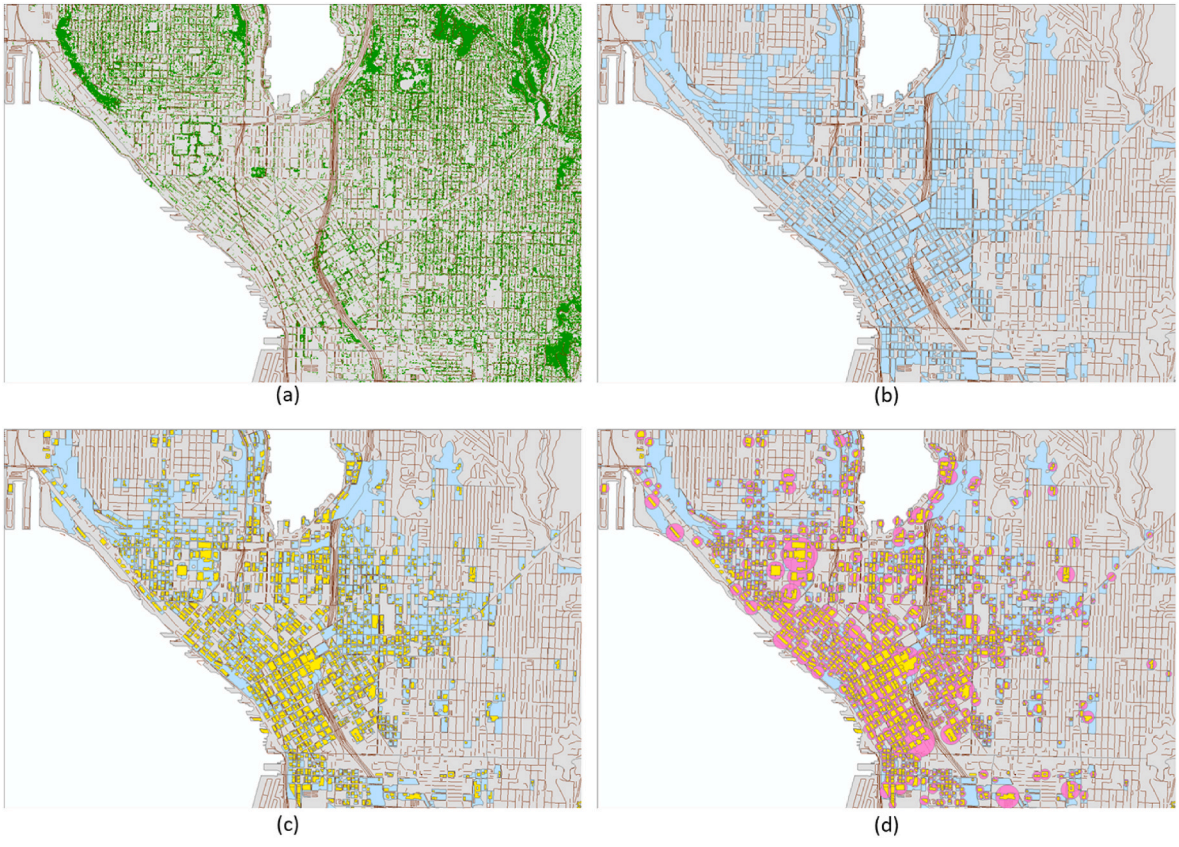

Fig. 4. An example of the (a) green coverage (canopy), (b) city blocks, (c) building footprint datasets, and (d) shape form factors in the study area.

Table 2
Description of building geometry variables.

Variables	Description	Reference
R_{VH}	Building's vertical to horizontal ratio	Bueno et al. [52]
BF	Building facade area	Bueno et al. [52]
A_{block}	Block area	Bueno et al. [52]
R_{BC}	Building coverage ratio	Yu et al. [53]
BA	Building footprint area	Yu et al. [53]
F_{block}	Block shape factor	Ma et al. [50]
A_{circle}	The minimum circumscribed circle of a specific building	Ma et al. [50]

so that the precision of model prediction is determined to be higher [57]. The process of the prediction is presented in Algorithm 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{obs,i}\right)^2}{\sum_{i}^{n}\left(y_{obs,i} - \bar{y}_{obs}\right)^2} \tag{5}$$

$$MSE = \frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{obs,i}\right)^2}{n} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{pred,i} - y_{obs,i}\right)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred,i} - y_{obs,i}\right| \tag{8}$$

where y_{pred} represents the predicted value, y_{obs} denotes the actual value, and $\bar{y}_{obs}$ is the average value of the observed data at n observations.

Algorithm 1. Predicting procedures using LightGBM.

```
Input: X dataset from {x1 ... x14} and two targets y1 and y2.
for x1 to x14 do:
    Generating both training and testing datasets.
    Identifying the parameters according to GridSearch.
    Training model based on LightGBM
        preds ← models.predict(X)
        g ← loss(X,preds)
        sorted ← GetSortedIndices(abs(g))
        topSet ← sorted [1:topN]
        randSet ← RandomPick(sorted[topNLlen(X)], randN)
        usedSet ← topSet + randSet
        models.append(newModel)
    Testing the Goodness-of-Fit for both y1 and y2.
end
Output: The prediction result of two targets y1 and y2
```

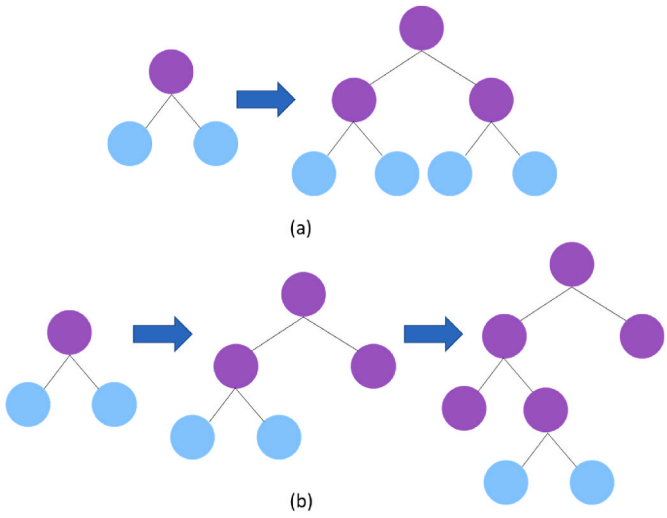

Fig. 5. Generation strategies of the tree: (a) Level-wise; (b) Leaf-wise.

$$g(x') = E[f(x)] + \sum_{i=1}^{K} E[f(x_i)]x_i' \tag{10}$$

where K is the number of the factor of the input factors, x_i' stands for a simplified factor i that can be reflected in the original factor x with a mapping function, $E[f(x_i)]$ is the corresponding Shapley value of the i-th factor, $f(x)$ represents the original model, and the mean value of the $f(x)$ indicates the predicted output $E[f(x)]$. The value of the simplified factor is 0 or 1, which means for "presence" or "absence", respectively. Specifically, when the $x_i' = 0$, then $E[f(x_i)] = 0$, representing that the factor has no impact on the final prediction. Thus, for a specific leaf node, where factor i is selected to split, all paths through this node collapse into two statuses, saying "present" or "absence".

Fig. 6 provides an example of the SHAP value of the summary plot and dependence plot. The SHAP value applied in the proposed model offers a reasonable estimation of the contribution of each feature. The fundamental concept behind this method is to utilize a simple model to forecast a complicated model that has been trained by the additive feature attribution method. Unlike the traditional bar chart only showing the feature importance, the summary plot in Fig. 6 (a) tells us which factors are most important and their range of effects over the dataset. The color allows us to match how changes in the value of a factor affect the change in output. The SHAP summary plot reveals a general overview of each factor, while a SHAP dependence plot shows how the model output varies by factor value. For Fig. 6 (a), the x-axis stands for SHAP value, and the y-axis has all the features. Each point on the chart is one SHAP value for a prediction and feature. Specifically, the

3.4. Model explanation

The SHapley Additive exPlanation (SHAP) method could be interpreted as an XAI method since it is established with a Shapley Value of Lloyd Stowell Shapley theoretical basis [58], aiming to distinguish from the conventional AI by providing a better understanding of the mechanism of the model in detail. This numerical approach quantifies the contributions of each factor to the overall outcome and explores the relationship between the inputs and outputs. It is a tree-based machine learning method, which can measure the influences of factors in the model and explain the model behavior. A liner function can be defined in Eq. (10) [59]:

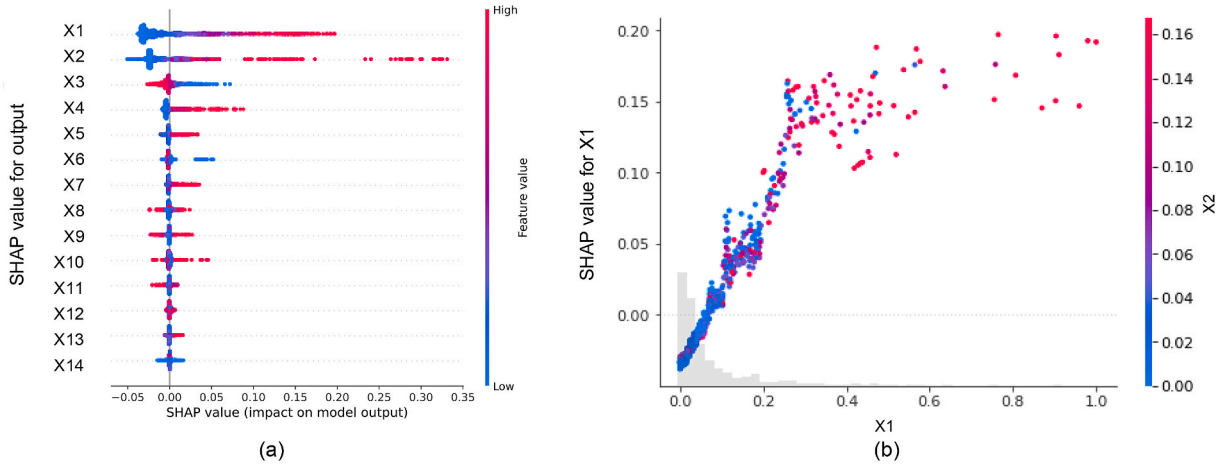

Fig. 6. An example of SHAP value in (a) summary plot and (b) dependence plot.

high level of the X1 has a high and positive impact on the output, where the "high" comes from the red color, and the "positive" impact is expressed on the x-axis. While, there is a negative relationship between X3 and the output, since the lower values of X3 lead to a higher SHAP value. Similarly, Fig. 6 (b) takes two input values, X1 and X2, as an example, it unveils that the SHAP value gradually grows up between 0.0 and 0.3 of factor X1 and stays almost unchanged when X1 increases from 0.5. Meanwhile, the upward trend in X1 colored by blue dots indicates that the value of X2 lies in the range between 0 and 0.04.

4. Case study

A case study is employed in this research to examine the proposed building energy evaluation framework and verify the accuracy of the LightGBM approach. Buildings' energy use and GHG emissions act as the two objectives for prediction in this study, where the most influential factors are identified by the SHAP algorithm. Findings are analyzed in the followed sections.

4.1. Background

This study was conducted based on the annual energy consumption of buildings and properties in Seattle. It has a land area of 217 km^2 and the population of Seattle has a number of 761,100, according to the Washington (WA) State Office of Financial Management. Fig. 7 (a) demonstrates the basic information of Seattle's surrounding zoning areas, where the single-family and industrial zoning make up more than 75% of Seattle's land area, and the majority of housing units (just over 60%) are currently found in other zones spread out over the city, including low-rise, midrise, high-rise, and neighborhood commercial zones. The dataset (https://www.seattle.gov/environment/climate-change/buildings-and-energy) consists of 3628 non-residential and multi-family buildings with 20,000 square feet or larger as shown in Fig. 7 (b). A total of 1523 data records of residential buildings remained after removing the outliers. Fig. 8 presents the Seattle city map with the individual building's energy use and GHG emissions. According to the Office of Sustainability and Environment in Seattle City, the average value for total site energy use and GHG emissions of the residential buildings within the study area are 2,814,333.25 kBtu and 51.14 $kgCO_2e$, respectively. However, the median values are 1,414,630.7 kBtu in energy use and 17.8 $kgCO_2e$ in GHG emissions. it is worth noticing that the mean values above the median value indicate higher overall energy consumption and very likely poor building energy performance.

In this study, we select the GHG emissions (in kgCO2e) and Site EUWN (weather-normalized site energy use in kBtu) from the dataset as the performance indicators of the building energy consumption. Weather-normalized site energy use adjust annual values based on that accounts for average and actual weather conditions. For instance, if a year had a higher than an average number of very cold days, buildings in that year would be expected to consume more gas than in an average

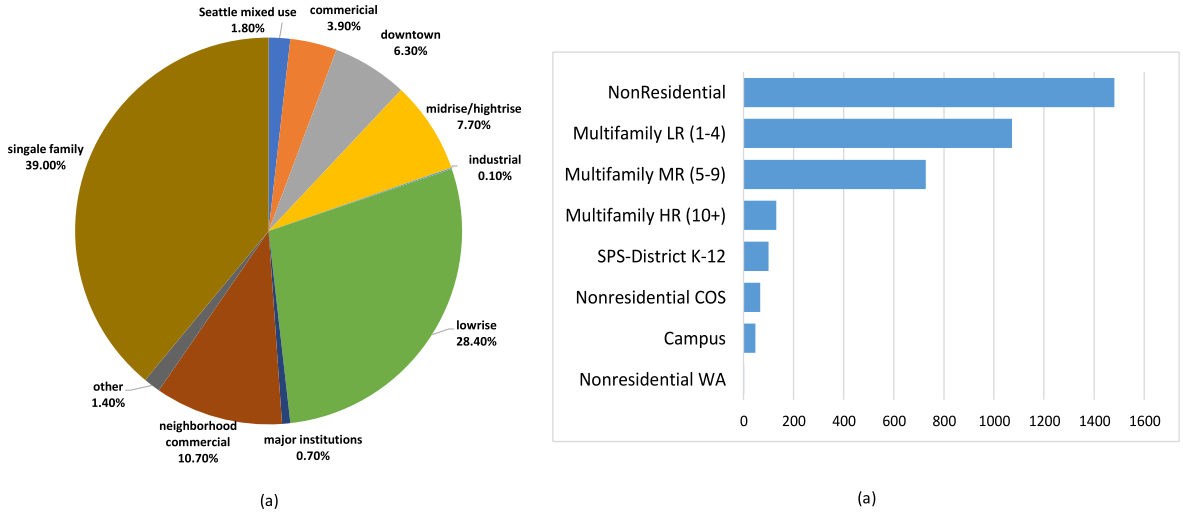

Fig. 7. Surrounding characteristics of the research object: (a) the zoning by dwelling units and (b) building types in Seattle city.

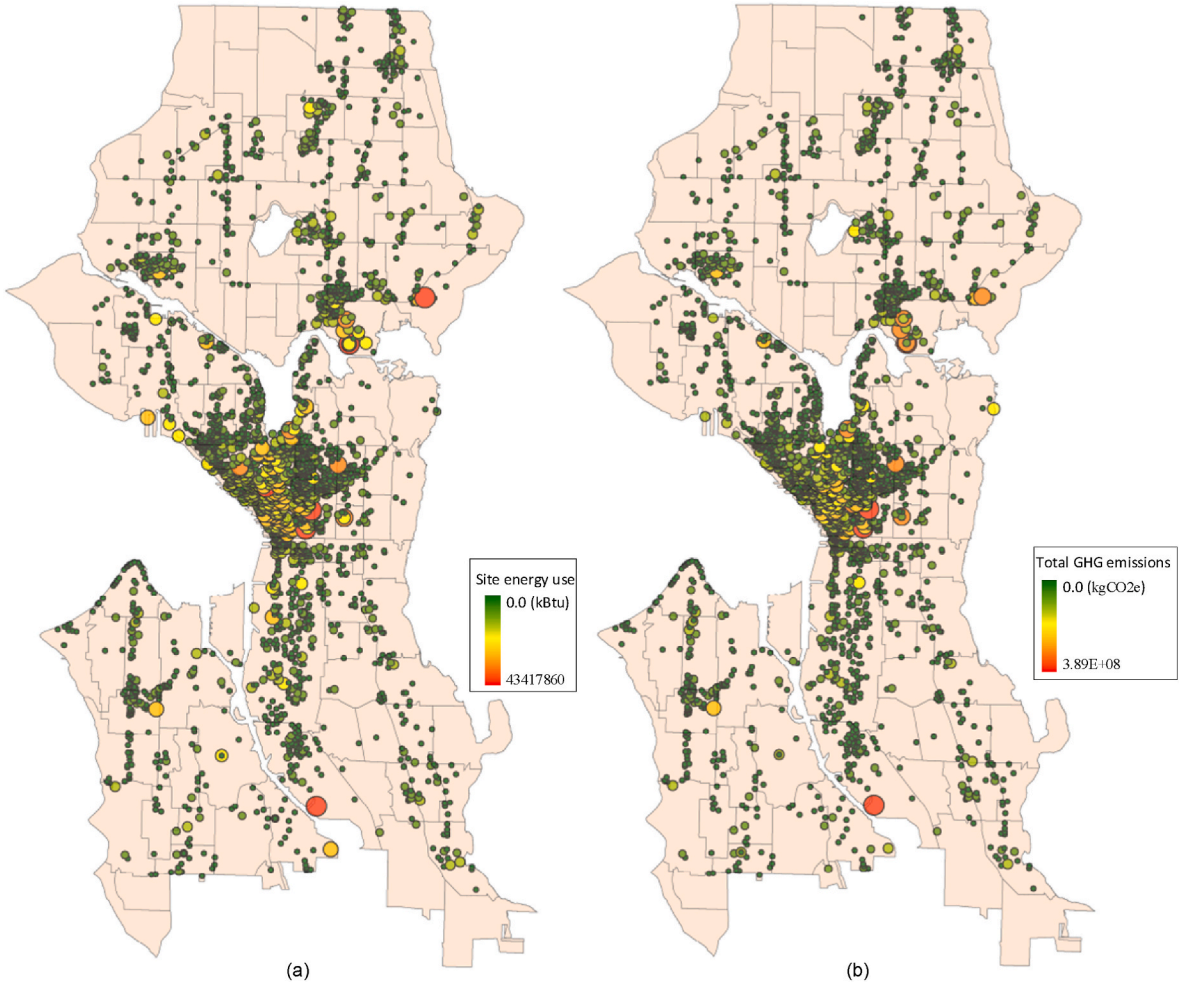

Fig. 8. Seattle city map of individual buildings with the measurement of (a) site energy use and (b) GHG emissions.

year. That year would also be expected to have a lower weather-normalized site energy use compared with its non-normalized consumption quantity. Table 3 provides a detailed description of identified factors and prediction objectives. The site EUWN is used as the benchmark for analyzing the energy performance among the buildings in the data record. A higher site EUWN record indicates a lower or poor efficiency in the building energy consumption. To further analyze the energy consumption of buildings, GHG emissions are also used as the indicator for each building to propose suitable solutions, such as policies or incentives, to reduce the environmental impacts caused by the properties.

4.2. Energy consumption and GHG emissions evaluation

To predict the energy performance and GHG emissions, a LightGBM regression model is constructed based on a LightGBM Python package with a total of 14 selected factors. Table 4 shows the basic information of the selected factors. A total of 1523 data records related to building energy consumption are used to construct the regression model. The dataset is split into 3 groups where the first 70% of the observed data are used as the training dataset (1066 data records) and 15% of the data as the validation dataset (228 data records) for developing the LightGBM model, following by the rest of 15% as the test dataset (228 data records) to analyze the prediction performance of the trained LightGBM model. Two separate regression models are established based on two different objectives, which are site EUWN and GHG emissions using the training dataset. The training dataset and testing dataset results are listed in Table 5.

The proposed LightGBM assessment model can achieve high accuracy of the prediction results, as the regressions for both energy use and GHG emissions are well-fitted as shown in Fig. 9 and Fig. 10. For the training data shown in the left column of Table 5, the values of R^2 for site EUWN and GHG emissions are 0.9269 and 0.8706, respectively. The values of MSE, RMSE, and MAE in site EUWN are 0.0006, 0.0254, and 0.0116, whereas the values for GHG emissions are 0.0008, 0.0278, and 0.0064, respectively. For the test data in the right column of Table 5, the value R^2 for site EUWN and GHG emissions is 0. 8608 and 0. 8261, respectively. The values of MSE, RMSE, and MAE for the model site EUWN are 0.0022, 0.0472, and 0.0248, respectively, whereas for the other target GHG emissions are 0.0013, 0.0363, and 0.0156, respectively.

4.3. Results analysis

To confirm that lightGBM has more accurate performance, different types of approaches are used to construct the regressors, and their effectiveness in predicting the building energy consumption is further evaluated in this study. The other three typical machine learning regression models, namely XGBoost, Random Forest (RF), and Support Vector Regression (SVR), are used to process the same dataset with LightGBM. Table 6 demonstrates and compiles the respective evaluation of regression models' accuracy for further comparison. To ensure the accuracy and robustness, the grid search method is adopted to optimize the kernel function parameter and penalty factor to ensure the model's

Table 3
Influential factors for the evaluation of building energy consumption and GHG emissions.

Features	Description	References
Building features		
X1. Natural gas	The annual volume of the usage of natural gas.	Wei et al. [48]
X2. Energy star score	A comprehensive snapshot of your building's energy performance, considering the building's physical assets, operations, and occupant behavior.	Dahlan et al. [47]
X4. Floor area per person	Floor space per person.	[60]
X3. Year of build	It reflects the age of a specific building.	Deb and Lee [46]
Building geometry		
X5. Total GFA	GFA is the total area of covered floor space measured between the centerline of party walls.	Ye et al. [45].
X6. Building height	Vertical distance above grade to the highest point of a flat roof.	Shareef [49]
X7. Shape form factor (F_{block})	The ratio between the building footprint and its minimum circumscribed circle.	Ma et al. [50]
X8. Vertical to horizontal ratio (R_{VH})	The ratio between its envelope area and its volume.	Hemsath and Bandhosseini [11]
X9. Length of building	Building's greatest horizontal distance or dimension.	Aldawoud [51]
X10. Width of the building	lesser of the two horizontal dimensions of a building or structure.	Aldawoud [51]
Urban morphology		
X11. Density of street	The ratio of the length of the country's total road network to the country's land area.	Ng et al. [39]
X12. Green coverage ratio	The ratio of the urban greenery to a certain area.	Fahmy et al. [41]; Ouyang et al. [42]
X13. Average number of buildings	The average number of a building within a certain area.	Yang and Wang [43]
X14. Building converge ratio (R_{BC})	The ratio of the building area is divided by the land (site) area.	[43,44]

accuracy and robustness, where the parameter, such as learning_rate and max_depth, are given in Table 6, Fig. 11, and Fig. 12 demonstrate the different results of the models for both EUWN and GHG emissions predictions. Detailed results are analyzed as follows.

(1) The comparison results prove that LightGBM is the most accurate model for predicting from Table 6, and the results from the other models are acceptable except for SVR. The average value of R^2 for the two targets predicted by LightGBM is 0.8435 ($R^2 = \frac{0.8608+0.8261}{2} = 0.8435$), which shows a growth rate of 4.67%, 7.75%, and 58.75% compared with the XGBoosting, RF, and SVR, respectively. Specifically, for site EUWN, the result of the LightGBM regressor has R^2 of 0.8608, which has a 5.79% improvement over XGBoosting, 8.15% over RF, and 16.91% over SVR as shown in Table 6. Also, for GHG emissions, the value of R^2 of LightGBM is 0.8261, which is 3.53%, 7.34%, and 153.17% higher than in XGBoosting, RF, and SVR, respectively. The other indicators, such as MSE, RMSE, and MAE, evaluated by LightGBM are 0.0022, 0.0472, 0.0248 and 0.0013, 0,0363, and 0.0156 for site EUWN and GHG emissions, respectively. It is followed by Random Forest Regressor, which demonstrates R^2, MSE, RMSE and MAE of 0.7959, 0.0016, 0.0403, 0.0168 and 0.7696, 0.0016, 0.0402, 0.0066 for the two evaluation targets. For the SVR model, though it has slightly higher accuracy in predicting site EUWN, the prediction of the GHG emissions has poor accuracy.

(2) Natural gas and total GFA are identified as the most common essential features in both site EUWN and GHG emissions, as they are ranked the top three as shown in Fig. 13. To identify the contribution of these 14 factors to energy consumption and GHG emissions, this study employed the SHAP value to analyze their importance. Fig. 13 shows the visualization results of the SHAP values, where each row represents a parameter feature with the SHAP value on the horizontal axis, and each point represents a data sample. The left parameter features in Fig. 13 are resorted based on the SHAP value. It is mentioned that the larger the SHAP value means the greater the impact of each feature on the target (i.e., the larger the sensitivity) [61]. For the energy consumption inFig. 13 (a), the top three important factors, including total GFA, natural gas, and energy star score are identified as the variables that are significantly associated with the building's energy consumption.

This result indicates that the larger the GFA is, the more energy consumed by buildings, consistent with the finding of previous studies conducted by Park et al. [62]. The natural gas and energy star score have

Table 5
Evaluation results for the regression models with training and test datasets.

Model	Training dataset		Test dataset	
	Site EUWN	GHG emissions	Site EUWN	GHG emissions
R^2	0.9269	0.8706	0.8608	0.8261
MSE	0.0006	0.0008	0.0022	0.0013
RMSE	0.0254	0.0278	0.0472	0.0363
MAE	0.0116	0.0064	0.0248	0.0156

Table 4
Selected factors and objectives for model prediction.

Factor	Unit	Mean	median	Min	Max	SD
YearBuilt	Year	1976.1390	1986	1900	2019	33.7209
Height	stories	22.1487	19.19	6.31	127.87	11.7966
Total GFA	sf	72464.74	41126	20000	678165	84773.95
Energy star score	score	78.2716	78.27	2	100	19.1568
Natural gas	kBtu	7379.0280	1373	0	186136	14851.06
Ave_BD	ratio	19.5903	16.818	6.745	127.87	10.94445
VtoH	numerical	1.5032	1.3509	0.0016	24.84372	1.088615
length	km	0.0503	0.0458	0.0099	0.1665	0.0225
width	km	0.0362	0.0322	0.0068	0.1556	0.0189
shape_form	numerical	32245.2	17656.99	1330.2750	391717.5	41648.87
BDcov	numerical	0.00921	0.0044	0.0002	0.143867	0.013805
GFA/pax	numerical	15.9925	8.9992	2.7373	263.1334	21.1137
DS_street/ft2	numerical	0.6136	0.4267	0.0115	9.7255	0.706319
DG_tree/sf2	numerical	10.3592	5.5579	0.0645	140.0927	14.8032
GHG emissions(kgCO2e)	kgCO2e	8268041	669992.8	0	3.89E+08	29423479
Site EUWN(kBtu)	kBtu	2694813	1388016	0	43417860	3762804

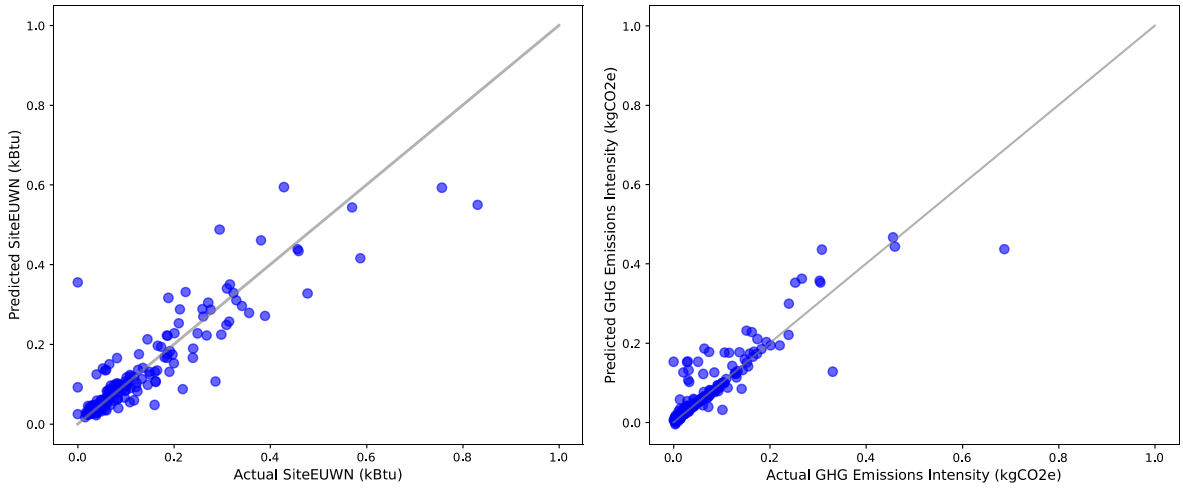

Fig. 9. Scatterplot of prediction results of LightGBM regression model for (a) site EUWN and (b) GHG emissions.

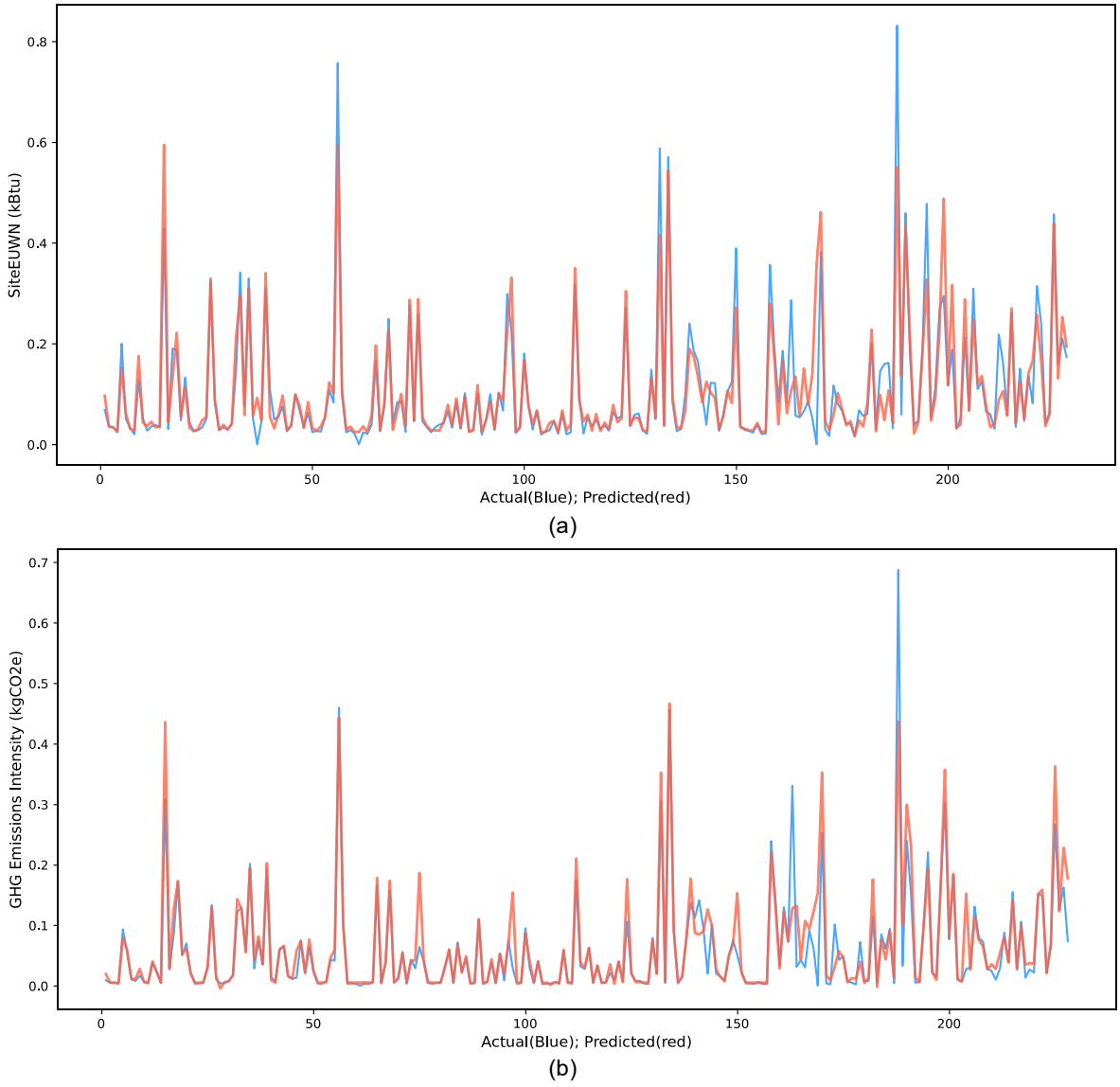

Fig. 10. Line plots of observed and predicted results of LightGBM regression model for (a) site EUWN and (b) GHG emissions.

Table 6
Comparison with different regressor models.

Model	Target	R^2	% changes in R^2	MSE	RMSE	MAE
LightGBM*	Site EUWN	0.8608		0.0022	0.0472	0.0248
	GHG emissions	0.8261		0.0013	0.0363	0.0156
XGBoost	Site EUWN	0.8137	5.79%	0.0006	0.0248	0.0123
	GHG emissions	0.7979	3.53%	0.0003	0.0128	0.0038
RF	Site EUWN	0.7959	8.15%	0.0016	0.0403	0.0168
	GHG emissions	0.7696	7.34%	0.0016	0.0402	0.0066
SVR	Site EUWN	0.7363	16.91%	0.0041	0.0640	0.0433
	GHG emissions	0.3263	153.17%	0.0050	0.0707	0.0627

Note*: learning_rate = 0.006, max_depth = 4, n_estimators = 900, num_leaves = 10.

opposite effects on site EUWN, where the more natural gas used by the household, the more energy consumption, which proves the conclusions reported by Copiello and Gabrielli [63]; whereas the higher energy star score indicates the less energy consumed. Moreover, the shape form and average building density have a relatively high impact on site EUWN. For the GHG emissions shown in Fig. 13 (b), it is observed that natural gas is the most sensitive factor to the GHG emissions among the 14 characteristics factors, followed by shape form, total GFA, year of built, and energy star score.

Three additional scenarios are studied to demonstrate the significance of building geometry and urban morphology for further investigating the impact of each factor on energy performance. To be specific, Scenario #1 is the original dataset that considers all features, including the building features, building geometry, and urban morphology. From Scenario #2 to #4, we remove each one of the impact aspects from the original dataset (Scenario #1). For example, Scenario #2 predicts the energy performance based on building features and urban morphology, while Scenario #3 is conducted relying on the building features and building geometry. Scenario #4 only considers the building features as the feature for evaluation. Table 7 lists the results of each scenario and Fig. 14 compares the prediction results. Detailed results are analyzed as follows.

(1) The comparison results indicate that impact of both building geometry and urban morphology on building energy consumption and GHG emissions cannot be ignored. For example, the values of R^2 in Scenario #1 are 0.8608 for site EUWN and 0.8261 for GHG emissions, where the average R^2 (R^2 = 0.8435) with

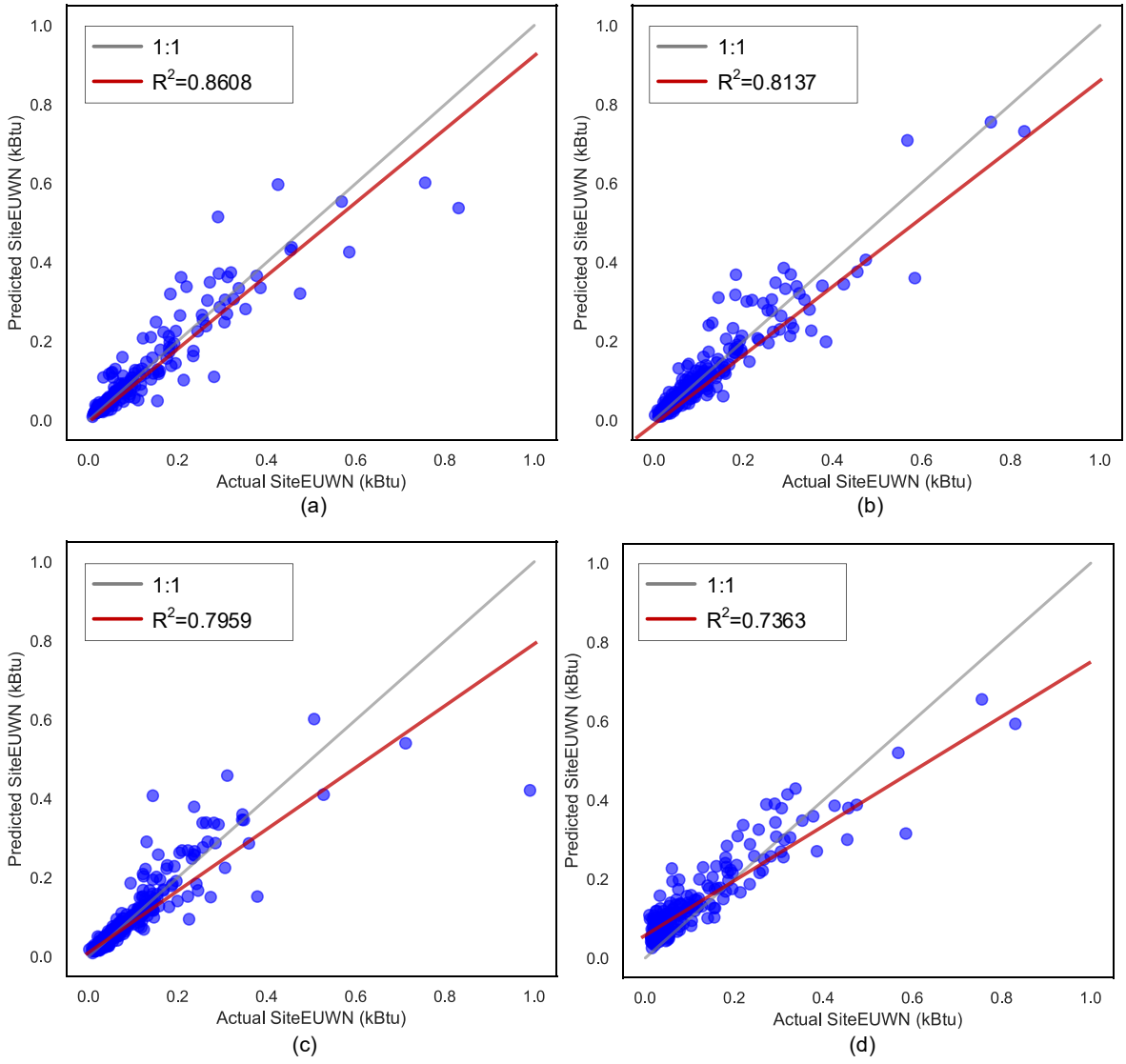

Fig. 11. Prediction results of the site EUWN from (a) LightGBM, (b) XGBoost, (c) RF, and (d) SVR models.

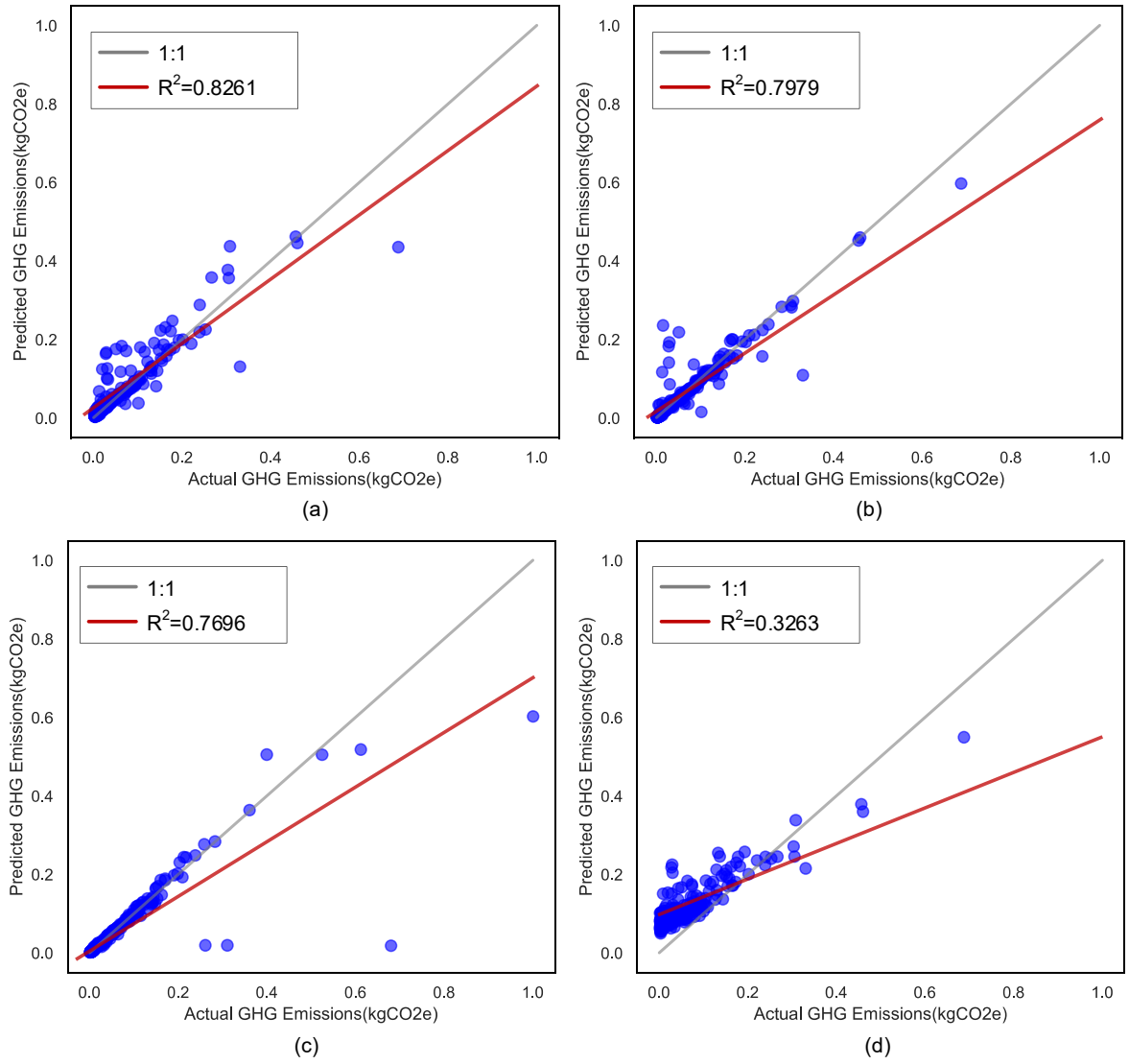

Fig. 12. Prediction results of the GHG emissions from (a) LightGBM, (b) XGBoost, (c) RF, and (d) SVR models.

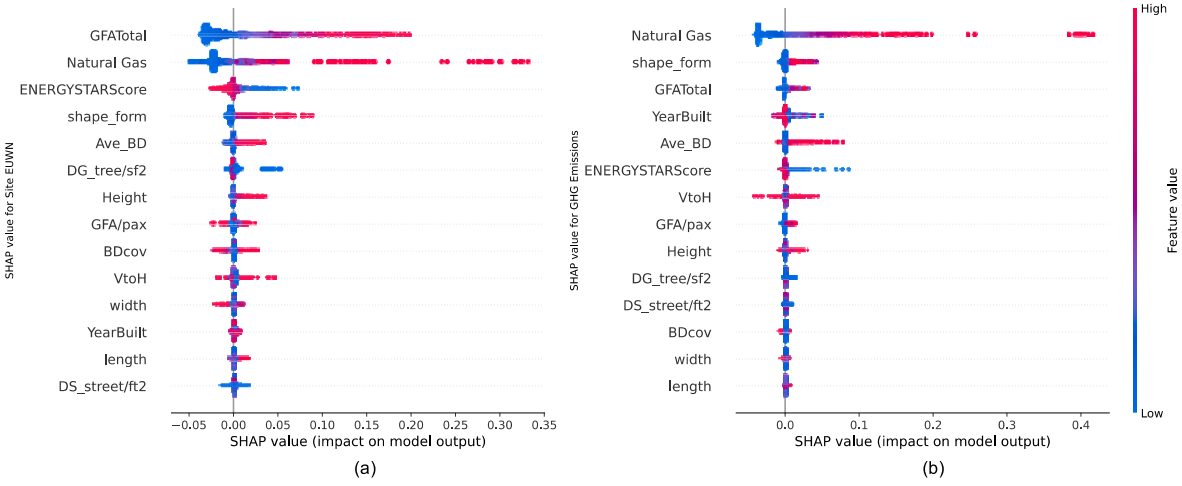

Fig. 13. SHAP value for (a) site EUWN and (b) GHG emissions.

Table 7
Comparisons of evolution on building energy performance in different scenarios.

Scenario	Model	R^2	MSE	RMSE	MAE	Percentage of average fluctuation
Scenario #1- all features	Site EUWN	0.8608	0.0022	0.0472	0.0248	/
	GHG emission	0.8261	0.0013	0.0363	0.0156	
Scenario #2-remove building geometry	Site EUWN	0.7615	0.0038	0.0617	0.3539	9.01%
	GHG emission	0.7734	0.0017	0.0414	0.0205	
Scenario #3- remove urban morphology	Site EUWN	0.7651	0.0038	0.0613	0.0321	13.18%
	GHG emission	0.6994	0.0023	0.0477	0.0240	
Scenario #4- remove both building geometry and urban morphology	Site EUWN	0.5679	0.0069	0.0831	0.0536	33.46%
	GHG emission	0.5546	0.0034	0.0581	0.0335	

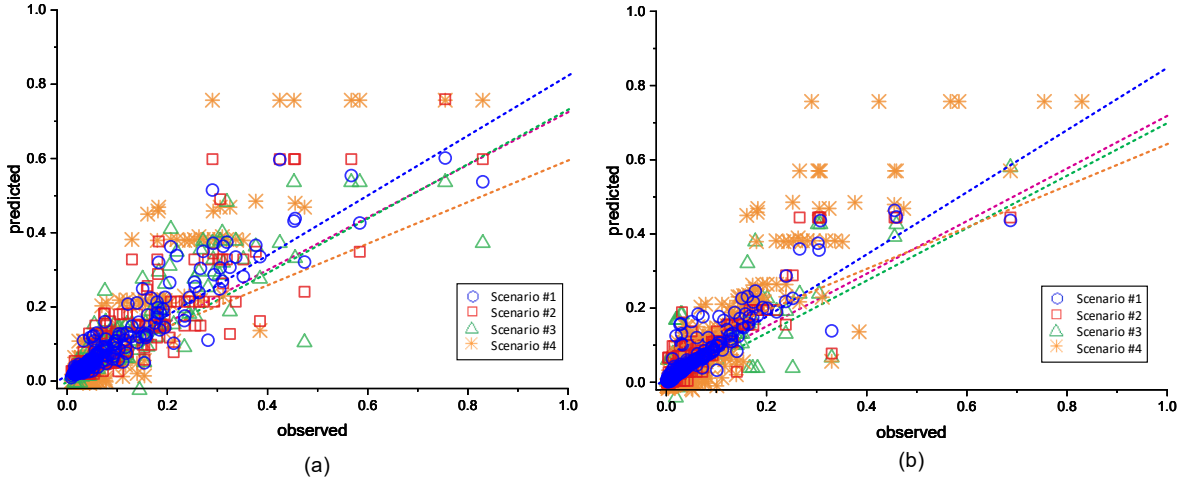

Fig. 14. Prediction results of four scenarios in (a) site EUWN and (b) GHG emissions.

taking all features into consideration is higher than Scenario #4 by 33.46%. However, the results of Scenarios #2 and #3, lacking building geometry and urban morphology features, respectively, are both below 0.8. Taking Scenario #2 as an example, the average value of R^2 is 0.7675, where the values for site EUWN (R^2 = 0.7615) and GHG emissions (R^2 = 0.7734) decrease by 11.54% and 6.38%, respectively, compared to the Scenario #1. For Scenario #3, the mean value of R^2 is 0.7322, which reduces by 13.23% on averagely compared with the original dataset, where the value of R^2 decreases by 11.12% in site EUWN and 15.34% in GHG emissions.

(2) Furthermore, the feature of urban morphology has a more significant effect on GHG emissions than site EUWN. Compared with Scenario #2 and #3, both building geometry and urban morphology have the same effect on the site EUWN, since the values of R^2 reduce to 0.7615 and 0.7651 in Scenarios #2 and #3, respectively. Nevertheless, the R^2 of GHG emissions cuts down to 0.6994 (15.34%) without analyzing the urban morphology, but it slightly decreases to 0.7734 (6.38%) without considering the building geometry feature. It reflects that GHG emissions are more sensitive to urban morphology, while the site EUWN has been equally affected by building geometry and urban morphology. Moreover, the result confirms that the urban morphology features cannot be ignored in analyzing the GHG emissions.

5. Discussions

To further explore the interrelationships between the pairs of input variables and their contributions to energy conservation and GHG

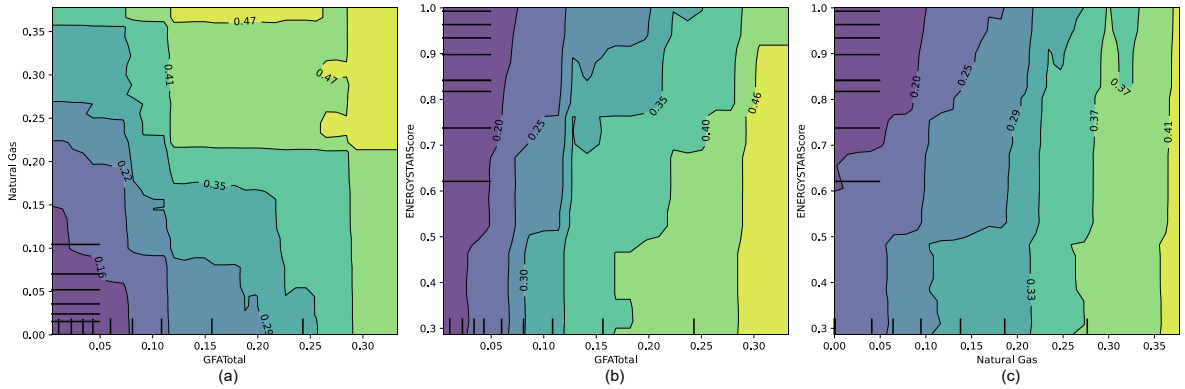

Fig. 15. 2D dependence plot of site EUWN with the top three factors in accordance with their largest negative average SHAP value under global scale analysis: (a) total GFA-natural gas; (b) energy star score-natural gas; and (c) energy stars core-total GFA.

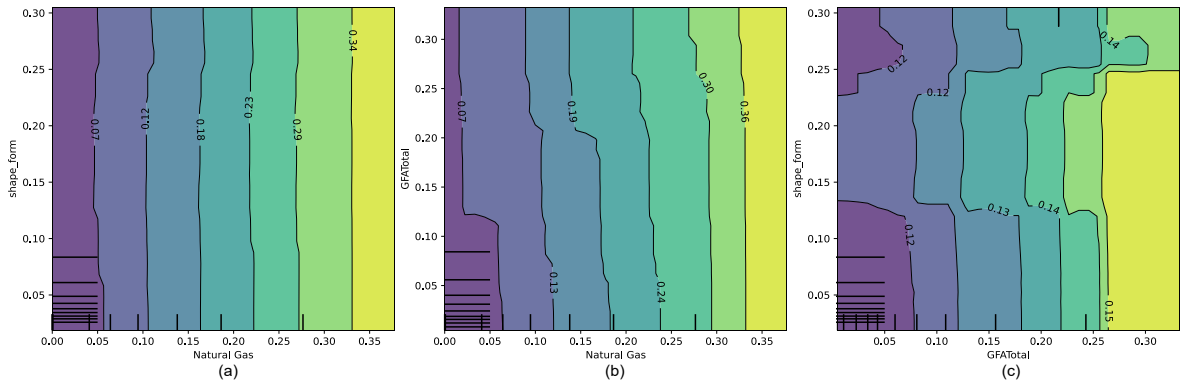

Fig. 16. 2D dependence plot of GHG emissions with top three factors in accordance with their largest negative average SHAP value under global scale analysis: (a) total GFA-shape form; (b) total GFA-natural gas; and (c) shape form-natural gas.

emissions reduction, a partial dependence plot (PDP) is illustrated in Fig. 15 and Fig. 16 to reveal interactions among the top three factors. All the values in the PDP figures are normalized values, and the contour line inside represents the value of the site EUWN and GHG emissions. For instance, the most significant variables, such as total GFA and natural gas, are combined in Fig. 15 (a) where the normalized values of 0.07 in total GFA and 0.15 in natural gas can best minimize the site EUWN in 0.16 during the model testing stage. It is followed by the energy star score and the total GFA (shown in Fig. 15 (b)), which can be interpreted that the combined normalized value of 0.29–1.0 for the energy star score and 0.06 for total GFA can lower the value of site EUWN of 0.2 approximately. For the interaction between the energy star score and natural gas presented in Fig. 15 (c), if the natural gas is below the normalized value of 0.12 and the energy star score above the value of 0.6, the total site EUWN could be minimized to 0.2.

Similarly, the interactions between the top three influential factors of GHG emissions are demonstrated in Fig. 16. Specifically, with the normalized range of 0–0.65 for natural gas and 0–0.31 for the shape form factor, the GHG emissions could be reduced to its minimal value of 0.07 as shown in Fig. 16 (a). Fig. 16 (b) indicates that the combination of the normalized range of 0–0.70 and 0–0.33 for natural gas and total GFA, respectively, would result in the lowest possible GHG emissions at 0.07. For the total GFA and shape form factor in Fig. 16 (c), the minimal value of the GHG emissions could be 0.12 under the situation that the normalized range of 0–0.077 in total GFA and 0–0.13 and 0.23–0.31 in shape form factor. Although the PDP displays interactions among the most important factors, more details are still underlying behind the results. For example, the distribution of the effects between factors and how is the effect of having a certain value constant or how it varies a lot depending on the values of other factors.

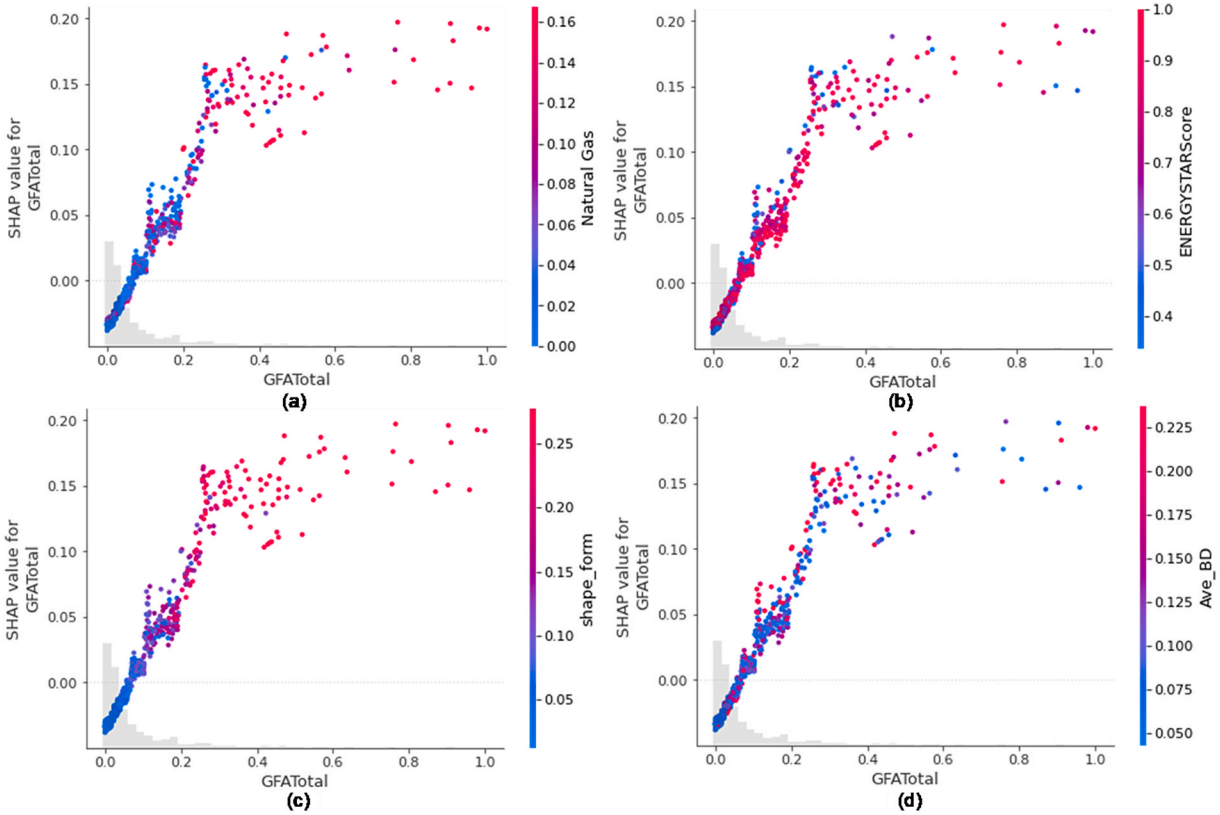

Fig. 17. Interaction effects of total GFA between the other top four factors: (a) natural gas, (b) energy star score, (c) shape form, and (d) average building density in site EUWN.

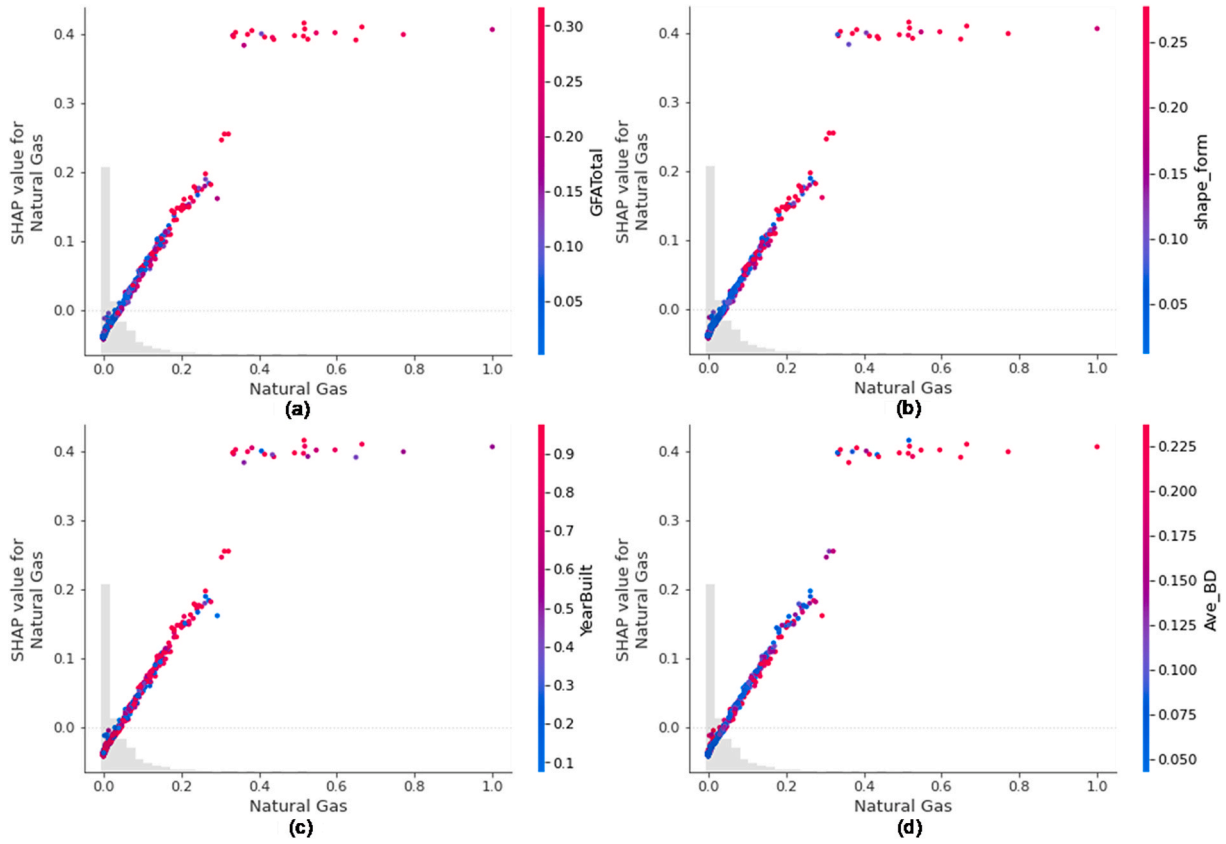

Fig. 18. Interaction effects of natural gas between the other top four factors::(a) total GFA, (b) shape form, (c) year of build, and (d) average building density in GHG emissions.

To complement extra details to the aforementioned PDP figures, a dependence contribution plot is provided as shown in Fig. 17 and Fig. 18, in which Figs. 17 and 18 illustrate how the other top influential factors affect the two objectives. As the total GFA and natural gas are identified as the most influential factor in site EUWN and GHG emissions, respectively, Fig. 17 fixes the x-axis as total GFA to illustrate the relationship of total GFA with the other top five important factors, and Fig. 18 designates the x-axis as natural gas to explore the interaction with the others. Meanwhile, the left y-axis represents how much of the contribution of that feature's value to the model prediction outputs. The right y-axis with the color bar from red to blue indicates the value of the influential factors. For site EUWN, the positive interaction effect can be found between total GFA-natural gas, total GFA-shape form, and total GFA-average building density, while the total GFA has a negative effect on energy star score. Fig. 17 illustrates the relationship of total GFA with the other top four important factors. Taking the interaction between total GFA and natural gas shown in Fig. 17 (a) as an example, with the increasing value of the total GFA, the value of natural gas becomes large, which indicates that there is a positive interaction effect between the natural gas and total GFA. In other words, if the building with large GFA, the total GFA-natural gas interaction effect increases as natural gas increases. Similarly, the interaction effects of total GFA-shape form and total GFA-average building density will improve with the larger natural gas consumption. By contrast, the interaction effect between the total GFA and energy star score has an opposite direction, where if the building with same natural gas consumption, the site EUWN will increase as the lower energy star score.

For GHG emissions in Fig. 18, it discusses the interaction effect of natural gas with the other four factors. A consistency tendency can be found that if the value of the influential factor is large enough, the natural gas will increase in the range of 0.4–0.6. To be specific, Fig. 18 (a) elaborates on the interaction between natural gas and total GFA and how much knowing the relationship changes the output of the model for GHG emissions. It is obverse to see that once the value of natural gas exceeds 0.38 (from the x-axis), the total GFA is stable at 0.3 (the red dot as shown in the color bar) and the SHAP value fluctuates between 0.4. While, for the natural gas with the range from 0 to 0.3, the value of total GFA climes (from blue and red dots) and the SHAP value for natural gas grows up to 0.2. Similar trends could be found in natural gas-shape form and natural gas-year built in Fig. 18 (b) and (c), respectively, and there is a slight difference in the value of the color bar. For instance, in Fig. 18 (c), most red dots representing the value of building year above 0.8 can be found when the value of natural gas is beyond 0.2. However, Fig. 18 (d) involving natural gas and average building density shows that the influence of the average building density on the GHG emissions is not significant in different natural gas, since the value of the building density is randomly located in the wide range of natural gas. This is to say, with the increasing value of the natural gas, the impact on GHG emissions is also getting server, which indicates natural gas has a positive effect on the GHG emission and interaction with the other top four factors except for the average building density.

6. Conclusions and future works

Energy conservation and GHG emissions reduction remain the main objectives for the building sector. Despite many studies focused on improving the energy efficiency of buildings, few of them consider the effects of building and urban morphology. Moreover, even though building geometry influences are discussed in some previous studies, they either ignore the urban morphology or adapt traditional statistical methods. Therefore, the LightGBM approach is applied to test the proposed assessment framework, and the XAI model demonstrated by SHAP is employed to better explain how urban morphology affects building energy performance and GHG emissions. Results confirm the effective

predictability of LightGBM by showing acceptable performance in accuracy indicators such as R^2, MSE, RMSE, and MAE. Also, the building geometry and urban morphology are proven as important features for analyzing energy consumption and GHG emissions. Moreover, this study contributes to urban planning in a similar scaled city by providing references for building owners and designers to make informed decisions, in order to promote suitable strategies in building energy efficiency improvement according to different city features, such as the climatic, building types, and the environments.

The case study in Seattle, U.S., is used to apply the proposed LightGBM method and to examine the potentially influential factors of energy use and GHG emissions. The main findings are (1) Urban morphology and building geometry are the two important features for forecasting building energy performance and GHG emissions. The value of R^2 taking these two features into account is 33.46% better than the initial situation; (2) Total GFA and natural gas are considered to be the most important factors affecting both site EUWN and GHG emissions, where there is a positive interaction between these two factors. It indicates that if the total GFA is enough large, the demand for natural gas would increase and result in a high site EUWN; (3) The proposed LightGBM approach, can achieve a higher accuracy of the result with the mean value of R^2 0.8435 for both site EUWN and GHG, which is better than the mean value of 0.8058 in XGBoost, 0.7828 in RF, and 0.5313 in SVR. Through the case study, the developed framework provides a comprehensive angle for estimating energy-related issues, resulting in a more accurate solution for urban planning. Meanwhile, it is expected to apply the proposed approach to other cities with similar scales, such as Los Angeles, Chicago, etc.

The study still has some limitations that are expected to be addressed in the future. Although some solutions for minimizing the site EUWN and GHG emissions are analyzed by the partial dependence method in the discussion part, specific optimization strategies need to be investigated to find the most ideal scenario to improve energy efficiency and reduce GHG emissions. Thus, multi-objective optimization (MOO) [64] could be a promising method to tackle energy-related issues. The identified important factors from the prediction model provide a clear direction for the optimization process, as these factors have a significant impact on the energy-saving target. In addition, as the evaluation focuses on the residential buildings, the other types of building with multiple targets could be investigated in the future to mitigate the negative effects of the building construction on the urban environment and sustainability.

Credit Author Statement

Yan Zhang: Writing – original draft, Methodology, Visualization, Investigation, Validation, Formal analysis.

Bak Koon Teoh: Methodology, Visualization, Formal analysis.

Limao Zhang: Conceptualization, Supervision, Methodology, Writing – review & editing, Funding acquisition.

Jiayu Chen: Methodology, Visualization

Maozhi Wu: Writing – review & editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (Grant No. 72271101). The 1st author is grateful to Nanyang Technological University, Singapore for providing the Ph.D. research scholarship.

References

[1] 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.
[2] Hirsh RF, Koomey JG. Electricity consumption and economic growth: a new relationship with significant consequences? Electr J 2015;28(9):72–84.
[3] Zheng Y, Weng Q. Modeling the effect of climate change on building energy demand in Los Angeles county by using a GIS-based high spatial-and temporal-resolution approach. Energy 2019;176:641–55.
[4] Delzendeh E, Wu S, Lee A, Zhou Y. The impact of occupants' behaviours on building energy analysis: a research review. Renew Sustain Energy Rev 2017;80:1061–71.
[5] Tibermacine I, Zemmouri N. Effects of building typology on energy consumption in hot and arid regions. Energy Proc 2017;139:664–9.
[6] Sun B, Luh PB, Jia Q-S, Jiang Z, Wang F, Song C. Building energy management: integrated control of active and passive heating, cooling, lighting, shading, and ventilation systems. IEEE Trans Autom Sci Eng 2012;10(3):588–602.
[7] Zhao H-x, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92. https://doi.org/10.1016/j.rser.2012.02.049.
[8] Wang L, Greenberg S. Window operation and impacts on building energy consumption. Energy Build 2015;92:313–21.
[9] Sonta A, Dougherty TR, Jain RK. Data-driven optimization of building layouts for energy efficiency. Energy Build 2021;238:110815. https://doi.org/10.1016/j.enbuild.2021.110815.
[10] Heydari A, Sadati SE, Gharib MR. Effects of different window configurations on energy consumption in building: optimization and economic analysis. J Build Eng 2021;35:102099. https://doi.org/10.1016/j.jobe.2020.102099.
[11] Hemsath TL, Bandhosseini KA. Sensitivity analysis evaluating basic building geometry's effect on energy use. Renew Energy 2015;76:526–38.
[12] Park J-H, Cho G-H. Examining the association between physical characteristics of green space and land surface temperature: a case study of Ulsan, Korea. Sustainability 2016;8(8):777.
[13] Srebric J, Heidarinejad M, Liu J. Building neighborhood emerging properties and their impacts on multi-scale modeling of building energy and airflows. Build Environ 2015;91:246–62. https://doi.org/10.1016/j.buildenv.2015.02.031.
[14] Peng LL, Jiang Z, Yang X, He Y, Xu T, Chen SS. Cooling effects of block-scale facade greening and their relationship with urban form. Build Environ 2020;169:106552.
[15] Strømann-Andersen J, Sattrup PA. The urban canyon and building energy use: urban density versus daylight and passive solar gains. Energy Build 2011;43(8).
[16] Güneralp B, Zhou Y, Ürge-Vorsatz D, Gupta M, Yu S, Patel PL, Fragkias M, Li X, Seto KC. Global scenarios of urban density and its impacts on building energy use through 2050. Proc Natl Acad Sci USA 2017;114(34):8945–50.
[17] Shashua-Bar L, Tsiros IX, Hoffman ME. A modeling study for evaluating passive cooling scenarios in urban streets with trees. Case study: Athens, Greece. Build Environ 2010;45(12):2798–807.
[18] Alznafer BM. The impact of neighbourhood geometries on outdoor thermal comfort and energy consumption from urban dwellings: a case study of the Riyadh city, the kingdom of Saudi Arabia. Cardiff University; 2014.
[19] Ma J, Cheng JC. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. Appl Energy 2016;183:182–92.
[20] Tong S, Wong NH, Jusuf SK, Tan CL, Wong HF, Ignatius M, Tan E. Study on correlation between air temperature and urban morphology parameters in built environment in northern China. Build Environ 2018;127:239–49.
[21] Zhang L, Li R. Impacts of green certification programs on energy consumption and GHG emissions in buildings: A spatial regression approach. In: 256. Energy Build; 2022, 111677. https://doi.org/10.1016/j.enbuild.2021.111677.
[22] Nemeth M, Borkin D, Michalconok G. The comparison of machine-learning methods XGBoost and LightGBM to predict energy development. Proc. Comput. Mthods Syst. Softw. 2019;1:208–15.
[23] Seyedzadeh S, Rahimian FP, Glesk I, Roper M. Machine learning for estimation of building energy consumption and performance: a review. Visualizat. Eng. 2018;6(1):1–20.
[24] Alam AG, Baek CI, Han H. Prediction and analysis of building energy efficiency using artificial neural network and design of experiments. Appl. Mech. Mater. 2016.
[25] Li Z, Dai J, Chen H, Lin B. An ANN-based fast building energy consumption prediction method for complex architectural form at the early design stage. Build Simulat 2019;12(4):665–81. https://doi.org/10.1007/s12273-019-0538-0.
[26] Ascione F, Bianco N, De Stasio C, Mauro GM, Vanoli GP. Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: a novel approach. Energy 2017;118:999–1017.
[27] Massana J, Pous C, Burgas L, Melendez J, Colomer J. Short-term load forecasting in a non-residential building contrasting models and attributes. Energy Build 2015;92:322–30.
[28] Li Q, Ren P, Meng Q. Prediction model of annual energy consumption of residential buildings. In: 2010 International conference on advances in energy engineering; 2010.

[29] Shao M, Wang X, Bu Z, Chen X, Wang Y. Prediction of energy consumption in hotel buildings via support vector machines. Sustain Cities Soc 2020;57:102128. https://doi.org/10.1016/j.scs.2020.102128.
[30] Liu Y, Chen H, Zhang L, Wu X, Wang X-j. Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: a case study in China. J Clean Prod 2020;272:122542. https://doi.org/10.1016/j.jclepro.2020.122542.
[31] Zeng A, Ho H, Yu Y. Prediction of building electricity usage using Gaussian Process Regression. J Build Eng 2020;28:101054.
[32] Yoon YR, Moon HJ. Energy consumption model with energy use factors of tenants in commercial buildings using Gaussian process regression. Energy Build 2018;168:215–24.
[33] Yuan T, Zhu N, Shi Y, Chang C, Yang K, Ding Y. Sample data selection method for improving the prediction accuracy of the heating energy consumption. Energy Build 2018;158:234–43. https://doi.org/10.1016/j.enbuild.2017.10.006.
[34] Parr T, Wilson JD. Partial dependence through stratification. Machine Learn. Appl. 2021;6:100146.
[35] Tso GK, Yau KK. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 2007;32(9):1761–8.
[36] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[37] Ahmad MW, Mourshed M, Rezgui Y. Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build 2017;147:77–89.
[38] Guo K, Zhang L. Adaptive multi-objective optimization for emergency evacuation at metro stations. Reliab Eng Syst Safe 2022;219:108210. https://doi.org/10.1016/j.ress.2021.108210.
[39] Ng E, Yuan C, Chen L, Ren C, Fung JCH. Improving the wind environment in high-density cities by understanding urban morphology and surface roughness: a study in Hong Kong. Landsc Urban Plann 2011;101(1):59–74. https://doi.org/10.1016/j.landurbplan.2011.01.004.
[40] Wei J, Ni Y, Zhang Y-J. The mitigation strategies for bottom environment of service-oriented public building from a micro-scale perspective: a case study in China. Energy 2020;205:118103. https://doi.org/10.1016/j.energy.2020.118103.
[41] Fahmy M, Mahdy M, Mahmoud S, Abdelalim M, Ezzeldin S, Attia S. Influence of urban canopy green coverage and future climate change scenarios on energy consumption of new sub-urban residential developments using coupled simulation techniques: a case study in Alexandria, Egypt. Energy Rep 2020;6:638–45. https://doi.org/10.1016/j.egyr.2019.09.042.
[42] Ouyang W, Morakinyo TE, Ren C, Ng E. The cooling efficiency of variable greenery coverage ratios in different urban densities: a study in a subtropical climate. Build Environ 2020;174:106772. https://doi.org/10.1016/j.buildenv.2020.106772.
[43] Yang Y, Wang P. Effects of building physics form on energy consumption for buildings. J Phys Conf 2022;2186:012008.
[44] Yu B, Liu H, Wu J, Hu Y, Zhang L. Automated derivation of urban building density information using airborne LiDAR data and object-based method. Landsc Urban Plann 2010;98(3):210–9. https://doi.org/10.1016/j.landurbplan.2010.08.004.
[45] Ye Z, Cheng K, Hsu S-C, Wei H-H, Cheung CM. Identifying critical building-oriented features in city-block-level building energy consumption: a data-driven machine learning approach. Appl Energy 2021;301:117453.
[46] Deb C, Lee SE. Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. Energy Build 2018;159:228–45. https://doi.org/10.1016/j.enbuild.2017.11.007.
[47] Dahlan NY, Mohamed H, Kamaluddin KA, Abd Rahman NM, Reimann G, Chia J, Ilham NI. Energy Star based benchmarking model for Malaysian Government hospitals - a qualitative and quantitative approach to assess energy performances. J Build Eng 2022;45:103460. https://doi.org/10.1016/j.jobe.2021.103460.
[48] Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, Han M, Zhao X. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018;82:1027–47. https://doi.org/10.1016/j.rser.2017.09.108.
[49] Shareef S. The impact of urban morphology and building's height diversity on energy consumption at urban scale. The case study of Dubai Building and Environment 2021;194:107675. https://doi.org/10.1016/j.buildenv.2021.107675.
[50] Ma R, Li X, Chen J. An elastic urban morpho-blocks (EUM) modeling method for urban building morphological analysis and feature clustering. Build Environ 2021;192:107646.
[51] Aldawoud A. The influence of the atrium geometry on the building energy performance. Energy Build 2013;57:1–5. https://doi.org/10.1016/j.enbuild.2012.10.038.
[52] Bueno B, Norford L, Hidalgo J, Pigeon G. The urban weather generator. J. Bulid. Perform. Simul. 2013;6(4):269–81 [Record #101 is using a reference type undefined in this output style.].
[53] Yu B, Liu H, Wu J, Hu Y, Zhang L. Automated derivation of urban building density information using airborne LiDAR data and object-based method. Landsc Urban Plann 2010;98(3–4):210–9.
[54] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.
[55] Gan M, Pan S, Chen Y, Cheng C, Pan H, Zhu X. Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia River. J Mar Sci Eng 2021;9(5):496.
[56] Moore DS, Kirkland S. The basic practice of statistics, vol. 2. New York: WH Freeman; 2007.
[57] Pan Y, Zhang L. Data-driven estimation of building energy consumption with multi-source heterogeneous data. Appl Energy 2020;268:114965.
[58] Park JH, Jo HS, Lee SH, Oh SW, Na MG. A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. Nucl Eng Technol 2022;54(4):1271–87. https://doi.org/10.1016/j.net.2021.10.024.
[59] Zhang L, Lin P. Multi-objective optimization for limiting tunnel-induced damages considering uncertainties. Reliab Eng Syst Saf 2021;216:107945. https://doi.org/10.1016/j.ress.2021.107945.
[60] Arehart JH, Pomponi F, D'Amico B, Srubar III WV. A new estimate of building floor space in North America. Environ Sci Technol 2021;55(8):5161–70.
[61] Dong X, Dong J, Zhou H, Sun J, Tao D. Automatic Chinese postal address block location using proximity descriptors and cooperative profit random forests. IEEE Trans Ind Electron 2017;65(5):4401–12.
[62] Park HS, Lee M, Kang H, Hong T, Jeong J. Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. Appl Energy 2016;173:225–37. https://doi.org/10.1016/j.apenergy.2016.04.035.
[63] Copiello S, Gabrielli L. Analysis of building energy consumption through panel data: the role played by the economic drivers. Energy Build 2017;145:130–43.
[64] Guo K, Zhang L. Multi-objective optimization for improved project management: Current status and future directions. Automat Constr 2022;139:104256. https://doi.org/10.1016/j.autcon.2022.104256.