Applied Energy 291 (2021) 116814

Contents lists available at ScienceDirect

Applied Energy

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

Space cooling energy usage prediction based on utility data for residential buildings using machine learning methods

Yanxiao Feng [a], Qiuhua Duan [a], Xi Chen [a], Sai Santosh Yakkali [b], Julian Wang [a,*]

[a] Department of Architectural Engineering, Pennsylvania State University, PA, USA
[b] Department of Civil and Architectural Engineering and Construction Management, University of Cincinnati, OH, USA

HIGHLIGHTS

- A user-friendly, infrastructure-free, and accurate model based on XGBoost.
- Hyperparameters tuning is applied to exploit optimal modeling performance.
- Interpreting cooling energy predictions and feature interactions by using SHAP.
- Surface to volume ratio estimated is beneficial for cooling energy use estimation.
- Degree hour data outperforms temperature for cooling energy estimation in XGBoost.

ARTICLE INFO

Keywords:
Cooling energy
Energy prediction
Utility data
Machine learning method
XGBoost model
Parameter tuning
Building characteristics
Weather features

ABSTRACT

The energy used for space cooling in residential buildings has a significant influence on household energy performance. This study aims to develop a user-friendly, infrastructure-free, and accurate prediction model based on large-scale utility datasets from anonymized volunteer homes located in three different climate zones in the US, along with the corresponding weather data and building information. Notably, several new weather- and building characteristics-related parameters were designed in the modeling procedure and tested to be useful for enhancing the model's prediction performance. A few regression techniques were examined and compared through hyperparameter optimization and k-fold cross-validation. Subsequently, a workflow was also described for how to implement the developed model. The research results showed that the eXtreme Gradient Boosting (XGBoost) model offered optimal performance, and the feature importance analysis also identified as well as ranked the key predictors to enhance the interpretability of this model. An R^2 value of around 97% was obtained with that model on the whole dataset, while an R^2 value of 92% was achieved with various subsets of the dataset through the cross-validation approach. The RMSE and RAE for this model were 0.294 and 0.153, respectively. The resultant model for predicting cooling energy consumption will facilitate homeowners better understanding their buildings' performance levels with minimum input information and without additional hardware installations, ultimately aiding their decision making related to energy-saving strategies.

1. Introduction

1.1. Background

The energy consumption of US households is increasing year by year, based on the residential energy consumption survey distributed by the US Energy Information Administration [1]. More specifically, the end-use sector of residential buildings comprised approximately 21% of the total US energy consumption in 2018. Space cooling, a major end-use in households, comprises averagely 14.7% of the annual home energy usage in the US [2]. Residential cooling energy demand can significantly affect aggregate energy consumption. Minor energy efficiency-based improvements could have a significant impact on the overall building energy used and potentially provide substantial energy cost savings. Studies also have indicated that residential energy use could be reduced through highly economical behaviors such as by developing encouraging cost-effective policies [3], taking effective intervention strategies [4], and providing energy use feedback to building occupants [5,6]. Thus, the collection of energy end-use data for

* Corresponding author at: 104 Engineering Unit A, University Park, PA 16801, USA.
E-mail address: julian.wang@psu.edu (J. Wang).

https://doi.org/10.1016/j.apenergy.2021.116814
Received 2 December 2020; Received in revised form 27 February 2021; Accepted 11 March 2021
Available online 1 April 2021

Nomenclature

Acronyms

ASHRAE	The American Society of Heating, Refrigerating and Air-Conditioning Engineers
CDD	Cooling degree days
CDH	Cooling degree hours
CV	Coefficient of variation
CV	Cross-validation
DBT	Dry-bulb temperature
HDD	Heating degree days
HDH	Heating degree hours
HHH	High humidity hours
HVAC	Heating, ventilation, and air conditioning
GHI	Global horizontal irradiance
LHH	Low humidity hours
NILM	Non-intrusive load monitoring
RAE	Relative absolute error
RH	Relative humidity
RMSE	Root mean squared error
SHAP	SHapley Additive exPlanations
SQL	Structured query language
S/V	Surface-area-to-volume
WBT	Wet-bulb temperature

Variables

A_{floor}	Total floor area [m^2]
E_{tc}	Total monthly energy use per unit area [kWh/sq·m]
h_{cd}	Cooling degree hours [°C·Hr]
h_{hd}	Heating degree hours [°C·Hr]
h_{hh}	High humidity hours [Hrs]
h_{lh}	Low humidity hours [Hrs]
T_b	Base temperature for the specific climate of the building's location [°C]
T_o	Average hourly dry bulb temperature [°C]
RH_b	Base humidity for the specific climate of the building's location [%]
RH_o	Average hourly relative humidity [%]
W_s	Monthly average wind speed [m/s]
Y_r	Age of house [Year]

use in evaluating and measuring current energy performance on the household level is essential.

1.2. Related work

The common approach adopted to resolve issues related to disaggregation and calculation of energy end-use using only total energy consumption data measured by a single meter is called non-intrusive load monitoring (NILM) which was primarily studied in the 1980s and 1990s [7]. NILM is economical and offers little complexity in terms of installation, and effectively and efficiently disaggregates total energy consumption into appliance-level energy use [8,9]. Earlier edge detection-based methods, e.g., Sultanem [10], Norford and Mabey [11], Laughman et al. [12] and Perez et al. [13], and subsequent machine learning-based algorithms such as hidden Markov models by researchers in [14,15], deep neural networks [16], and k-nearest neighbor (KNN) [17] mostly rely on datasets comprised of circuit-specific electricity usage to implement appliance-level energy monitoring. In other words, NILM doesn't mean hardware-free. Measurement of circuit-specific data for high-frequency current and voltage waveform still relies on specific monitoring hardware and requires certain levels of computation for further analysis. The Reference Energy Disaggregation Dataset is a representative data sample for NILM-based methods that has frequently been utilized to explore different calculation algorithms [18]. In addition to restrictions on data collection, it is also difficult to generalize inferred end-use results for convenient and practical application by building users.

Other widely used alternative statistical approaches include establishing quantitative correlations between cooling energy consumption and influential factors, and developing prediction models based on historical data obtained from utility bills. It is generally understood that both the climatic environment and building-specific information have a significant influence on a building's energy consumption. Earlier studies, e.g., by Sonderegger [19], Dhar et al. [20] and Dong et al. [21] have examined the substantial influence of weather-related parameters, utilizing regression methods and outdoor temperature and degree-days parameters for 12 months to predict a baseline of daily and monthly aggregated energy consumption for a single building, without addressing appliance-level energy consumption. Study in [20] developed a simplified Fourier series model to predict hourly cooling energy consumption for one year, considering the effects of outdoor temperature, humidity, and solar radiation using hourly monitored energy data from several buildings in the same city. Similar influential parameters were considered in research [21] using support vector machines (SVM) to develop and test a prediction model for total energy consumption, using utility bills for commercial buildings in a tropical region. The sample commercial buildings in these two studies were located in one climate region. Thus, further work is needed to verify the model's applicability to building sites with different climate features. The effect of the length of the measurement period on prediction accuracy was evaluated in [22]. The authors argued that prediction accuracy would be improved with energy measurement data collected over a longer time span.

The aforementioned statistical approaches have delved into the real-world problem of HVAC energy consumption, and all have reported acceptable levels of performance normally evaluated by goodness-of-fit indicators such as coefficient of determination (R^2), percentage error, coefficient of variation (CV) of root mean squared error (RMSE) and relative absolute error (RAE). The results provided by [21] had a CV value smaller than 3% and a percentage error less than 4%, while [23] obtained results with an average deviation of 3.4% and a prediction rate ranging from 94.8% to 98.5%. However, their limitations cannot be ignored. The datasets used in these studies are typically obtained from a very limited number of buildings and mostly in the commercial building sector. The weather feature selected in these studies is also limited to a particular climatic zone. More importantly, building-specific information such as size, geometry, etc., have not been taken into account in the modeling process. To address these issues, some studies such as [23,24] have incorporated building energy simulation programs to add building-specific information such as orientation and insulation thickness and used the output of simulated energy use as reference values for evaluating or enhancing the prediction accuracy of their models. An improved method using Gaussian Process Regression was proposed in [25] to model the electricity use of office buildings. Actual energy consumption data were used as the validation data to calculate accuracy. However, the research objective of these simulation-related studies mainly focuses on overall building energy assessment, optimization, or system diagnosis rather than generalizable models to decompose the whole building energy to end-use. In brief, these machine learning-based statistical methods are promising for predicting performance but still rely on sensor or monitor installation to form a large enough training dataset for modeling. Moreover, some weather-based energy disaggregation methods have provided infrastructure-free opportunities, but it remains challenging to achieve high accuracy and generalizability in models based on pure weather data. Although some research effort has involved

simulated energy use data, building-specific information and variety have not been fully engaged.

In addition to the NILM, weather-based, and simulation-incorporated methods described above, some studies have explored the utilization of more user-ended information such as utility bills for specific energy use prediction and then recommended and/or implemented energy savings strategies. Local governments and public utilities are now supporting grid modernization in collaboration with other influential parties by deploying technologies such as smart sensors, computers, and telecommunications, making utility bills more accessible [26]. Moreover, a home scoring methodology was proposed by the US Department of Energy in 2017, which evaluates a building's overall energy performance as compared to similar building types and provides cost-effective retrofits based on two influential factors [27]. Publicly shared large-scale datasets comprised of utility bills provide new opportunities for space cooling energy usage prediction and energy efficiency improvements that can be generally applied. Additional energy bill analysis methods have been explored to identify and simplify model input parameters while ensuring prediction accuracy. Researchers in [28] utilized an optimization algorithm to disaggregate energy bills into three groups of end uses based on the principles of electricity and cooling energy balance, considering input data including monthly electricity bills, building design data, weather conditions, and HVAC operation information. However, that model requires relatively detailed HVAC user input data on specific heating and cooling energy consumption. The authors of [29] developed a change-point linear regression model, using utility billing data and weather information to estimate gas and electricity usage for end uses such as space cooling and heating in residential buildings with radiant floor heating systems installed. The proposed method can accurately predict gas used for heating but had only average prediction accuracy for electricity consumption. [30] conducted a multiple linear regression analysis of commercial buildings' overall energy consumption in tropical regions, using utility bills and weather data; the authors argued that more significant variables were needed to improve estimation accuracy. However, in that work, due to the limited input information, issues related to parameter collection and identification for the models' prediction performance were identified, and the practical usefulness to homeowners was challenged. Briefly, utility bill-based methods offer new opportunities for energy disaggregation, but accurate monthly cooling energy use with limited user input has yet to be achieved.

1.3. Motivation and contribution

Although the statistical regression models have been applied in many studies as described above and obtained satisfactory prediction results, some limitations or challenges are identified in these studies. Firstly, sample buildings used in such modeling research were mostly situated in the same city and climate regions and locations which decrease the input data information quantity and complexity for model enhancement. Second, the weather data and building information were not fully considered and incorporated for the input predictors to facilitate the modeling procedure. Thus, the modeling process was complex in order to guarantee a satisfactory prediction accuracy, and the developed model is difficult to be implemented by homeowners with the complicated processed input information. Third, the interpretability of developed machine learning models in terms of the feature importance and interactions has received more attention in recent years but still has not been studied fully. Moreover, the performance of commonly used machine learning methods in estimation building cooling energy consumption has not been compared thoroughly in terms of their prediction accuracy and identifying key predictors. In this work, several research questions were of particular focus during the modeling processes to address the issues in previous studies, including how to process and yield more useful weather data from readily available weather databases, how the first principles of building physics should be combined, and what new (readily available) parameters might be incorporated to facilitate the modeling procedure and to simplify the manipulation process for homeowners with the minimum input information. We aim to apply machine learning modeling techniques and methods with parameters optimization to develop a user-friendly, infrastructure-free, and accurate prediction model for estimating cooling energy use based on readily available utility bills, weather data, and simple user input information (i.e. physical home address) to simplify the manipulation process for homeowners with the minimum input information.

Relative to previous studies, one major contribution of this work is that we intentionally bring building characteristics such as floor area, age, and surface-to-volume ratio into the modeling process, yielding a more generalizable model that can be applied to various residential buildings. Another novelty is that several new weather- and building characteristics-related parameters by combining the first principles of building physics have been proposed, processed, analyzed, examined, and determined to be effective at enhancing the model's performance regarding cooling energy use prediction. Also, different prediction models are compared in terms of their prediction power and computation complexity through hyperparameters tuning and k-fold cross-validation. More practically, sensitivity analysis by interpreting feature interactions is conducted for the input variables to identify the most influential variables in the models. From a broader scope, the social impact of this work is in its support of homeowners seeking to quantitatively understand their use of building heating and cooling energy, ultimately aiding their energy-saving-related decision making.

The remaining of this paper is organized as follows: Section 2 provides the modeling processes for disaggregating total energy consumption into space cooling energy use and introduces a few new parameters with underlying building physics principles. Section 3 presents the results of the different modeling methods. Section 4 discusses problems and limitations related to the input variables that affect the accuracy of the prediction, as well as implementation workflows of applying the established model to solve real-world home energy issues. Finally, conclusions are outlined in Section 5.

2. Method

2.1. Data collection

The volunteer buildings selected for model development were anonymized via exclusive IDs provided by the Pecan Street organization [31]. The buildings were located in the cities of Austin, Texas, Boulder, Colorado, and San Diego, California. These three cities represent hot-humid (Zone 2A), cold (Zone 5B), and hot-dry (Zone 3B) climates, respectively. Hourly utility data were collected and converted using structured query language (SQL) commands to obtain the hourly, daily, and monthly total energy consumption levels, as well as space cooling energy usage over the span of four years, from 2014 to 2017 [31]. After preliminary data processing such as extraction of the feature information, deletion of missing data, and removal of outliers, 391 houses located in the above three cities, a total of 6690 observations were used to conduct the monthly cooling energy usage per unit area modeling process. The distribution of the observations for each of these three cities is shown in Fig. 1.

The corresponding weather information for the three cities from 2004 to 2007 could also be accessed and retrieved from the same database provided by the Pecan Street organization through SQL commands [31]. The typical weather parameters, which were directly obtained, included the dry-bulb temperature (DBT), wet-bulb temperature (WBT), relative humidity (RH), and wind speed. The hourly weather data for these parameters were also converted to monthly data by averaging all the values in a month. The historical monthly solar radiation data for the three cities were obtained from the TMY3 database [32]. The monthly average DBT and global horizontal irradiance (GHI) for the three cities are shown in Fig. 2.

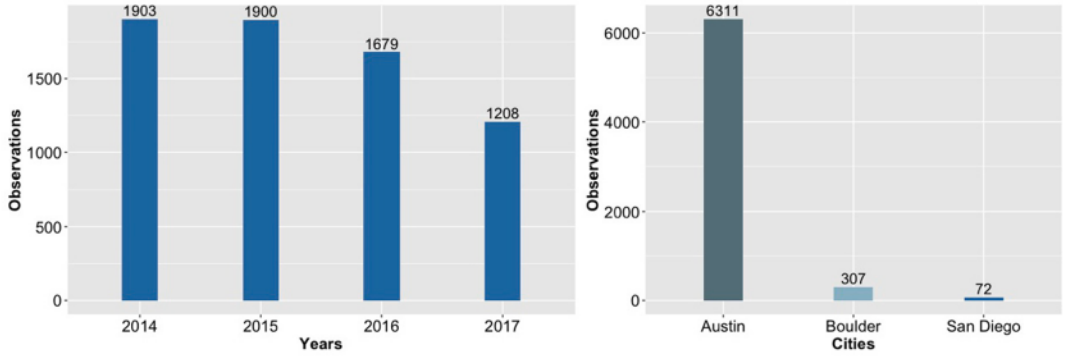

Fig. 1. Observation distribution for sample buildings.

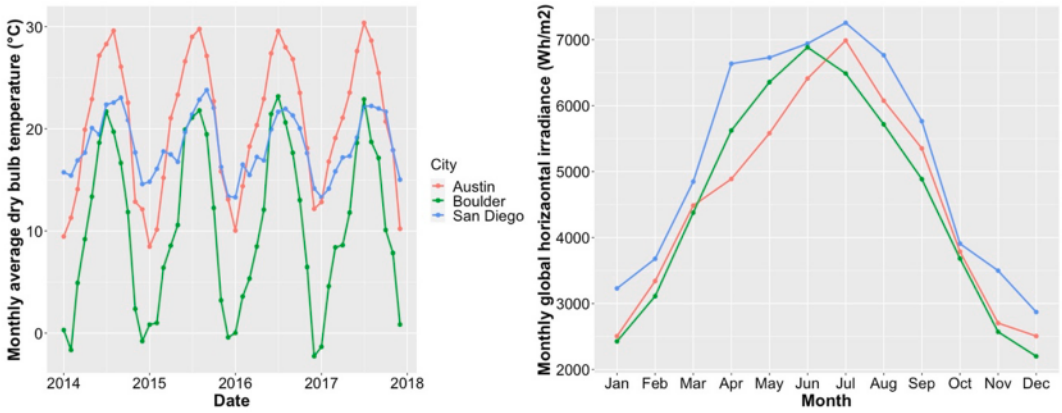

Fig. 2. Monthly temperature and GHI for the three cities.

In addition, the metadata downloaded provided general building characteristics such as the year built, building floor area, number of floors, and city location, which from the future application's perspective could also be retrieved from public property records based on home address input. The total energy consumption, as well as space cooling per square foot of the house, were also calculated to make figures comparable for homes of various dimensions.

2.2. Data pre-processing

2.2.1. Determination of summer cooling seasons

The monthly cooling energy usage per unit area for all houses located in the three cities are summarized along the left y-axis of Fig. 3. There is an apparent difference in monthly energy consumption. Although different energy usage patterns appear for each of the three cities, in hot weather seasons the buildings generally consumed much more

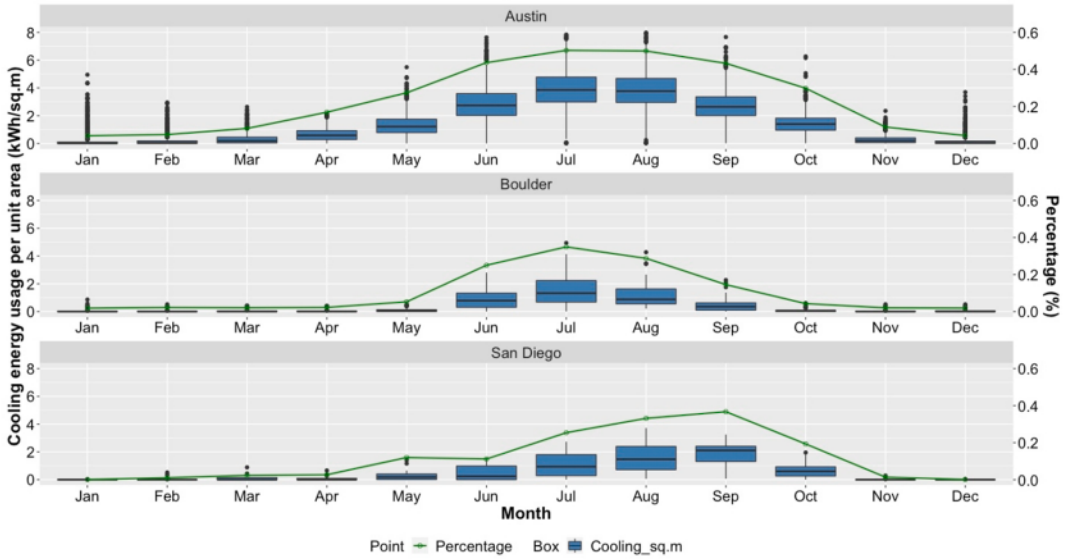

Fig. 3. Monthly distribution of cooling energy and percentage of manifest energy usage by month.

electricity than they did in cold weather seasons. To develop a robust and accurate energy estimation model that fits the bulk of the data, unqualified data that might affect the model's performance were identified, such as the minimum values in cold seasons and outliers. Austin had obvious hot season cooling energy usage from May to October, with peak values occurring in July and August. The maximum monthly cooling energy usage could be around 10.0 kWh/sq·m. Boulder and San Diego had less cooling energy demand compared to Austin. Similarly, the maximum cooling energy usage was around 4.0 kWh/sq·m for Boulder and 3.5 kWh/sq·m for San Diego.

To obtain a more specific cooling energy use dataset for cooling seasons determination, we analyzed the percentage of monthly cooling energy use as a proportion of monthly total energy use. In this work, we used 5% as the threshold to determine the cooling season or cooling months to ensure the reliability and validity of the measured data. In other words, the cooling energy use data were assumed to be 0 and not considered in the modeling procedure if the cooling energy percentages fell to 5% of monthly total energy use. The right y-axis in Fig. 3 illustrates this information. The percentages during certain months are very small or negligible, though the period has slight differences across the three cities. In particular, these very small numbers may contain errors induced by hard-to-detect or subtle reasons such as degradation of measurement devices or occasional user cooling needs and behaviors (e.g., indoor physical activities).

2.2.2. Heating and cooling degree hours calculations

Weather-related parameters usually focus on temperature and humidity. Temperature-related parameters, including DBT, WBT, and dew point are customarily included in this type of model. In most previous studies, the degree-day parameters of cooling degree days (CDD) and heating degree days (HDD) are typically considered and exhibited as significant influences on energy use for building heating and cooling purposes. We borrowed the analogous concept with higher-resolution data: cooling degree hours (CDH) and heating degree hours (HDH) that have been explored in some studies such as by researchers in [33,34] to present the distribution of hourly average temperatures and in [35,36] to estimate the energy usage of a building based on degree-hour method. This hourly-based calculation quantifies the cumulative value of the difference between the base temperature and hourly average temperature, rather than the average daily temperature considered in degree-day calculations. The motivation for this data processing was that some large temperature swings could be ignored in the calculation of CDD and HDD. For instance, similar CDD values can be yielded by two different weather situations that may drive very different space cooling energy use levels. Comparatively, the hourly-based calculation method for CDH and HDH does not merely utilize a database with the hourly resolution but also considers possible diurnal temperature variations hidden in the weather dataset. The CDH and HDH are calculated following Eqs. (1) and (2), respectively.

$$CDH = \sum_{n=1}^{N} (T_o - T_b)^+ \quad (1)$$

$$HDH = \sum_{n=1}^{N} (T_b - T_o)^+ \quad (2)$$

where CDH is the cooling degree hours, HDH is the heating degree hours, N is the number of hours in the defined period of analysis, T_b is the base temperature for the specific climate of the building's location, and T_o is the average hourly DBT.

Base temperature varies with the climate zone, and several different values have been defined in the ASHRAE Handbook 2017 [37]. The base temperatures for CDH and HDH calculation are commonly at the same value, or a larger base temperature for calculation of CDH than HDH is used, e.g., Papakostas and Kyriakis [34], Bolattürk [35], and Shi et al. [38]. The correlation coefficient was calculated to evaluate the influence of base temperature on the correlation between space cooling energy usage per unit area and CDH or HDH for the selected buildings. The results, as shown in Fig. 4, indicate their statistical relationship with the base temperatures for CDH and HDH, ranging from 15 °C (60°F) to 38 °C (100°F) and 4 °C (40°F) to 27 °C (80°F), respectively. For both the CDH and HDH calculations, the correlation coefficient has an approximate maximum value when the base temperature equals 23.8 °C (75°F). The selection of this base temperature maximizes the correlation between the degree-hours and cooling energy usage while complying with the practical norm and avoiding overlapping in the cooling and heating temperature ranges. The base temperature, defined as 23.8 °C to harmonize the degree hours calculations for CDH and HDH, is also included because there is no noticeable difference in variations of the correlation coefficients for cooling energy use. The monthly accumulated CDH and HDH calculated for the three cities from 2014 to 2017 are shown in Fig. 5. It is evident that the CDH in Austin is much larger than those in the other two cities, and the maximum is around 4200 in the cooling season. There is little difference between the CDH values for the other two cities. The maximum numbers are 1100 and 1400 for San Diego and Boulder, respectively. Also, the distributions of CDH and HDH values over the course of a year appear to follow similar trends for each city in those four years.

2.2.3. Low and high humidity hours calculations

Home space air conditioning energy use, especially cooling energy use, can be affected by high humidity conditions. This has been studied in several works, such as [39,40]. For air conditioning systems embedded with humidity sensors, external humidity levels may influence both heating and cooling energy use. Therefore, similar to the temperature parameter, low humidity hours (LHH) and high humidity hours (HHH) were defined in this study to increase the resolution and completion of the weather data. The LHH and HHH are represented by Eqs. (3) and (4), respectively.

$LHH = \sum_{n=1}^{N} (RH_b - RH_o)^+$ (3)

$HHH = \sum_{n=1}^{N} (RH_o - RH_b)^+$ (4)where LHH is the low humidity hours, HHH is the high humidity hours, N is the number of hours in the period of analysis, RH_b is the base humidity for the specific climate of the building's location, and RH_o is the average hourly RH.

LHH and HHH may both affect cooling energy usage, so the correlations between cooling energy usage and LHH and HHH with RH values ranging from 10% to 80% were examined, as shown in Fig. 6. Both LHH and HHH negatively correlated with space cooling energy consumption, and the maximum correlation occurred when the base RH was 30% for LHH computation. In sum, the statistical correlation of HHH with space cooling energy usage was relatively small. The coefficient slowly approached a maximum value when the base RH exceeded 80%. The base RH values of 30% and 80% were selected. The monthly accumulated humidity-related hours for the three cities are shown in Fig. 7, where it can be seen that the values for LHH and HHH had an unexpectedly patchy distribution pattern.

2.2.4. Surface-area-to-volume ratio estimation

In the domain of building physics, different types of research have been conducted on the effects of building geometry on energy use, involving numerous aspects of building geometry, such as surface-to-volume ratio [41], building orientation [42], building form [43], and dynamic envelope [44,45]. The surface-area-to-volume (S/V) ratio, a common measure of building compactness, has been found to significantly impact overall heat gains and losses through a building envelope. It is considered one of the building-specific variables most important to consider [41]. It is commonly accepted that a compact building shape introduces less heat gain and loss, and at the same time may interactively influence heat transfer through the envelope via complicated factors such as solar radiation, wind speed, and building orientation. The S/V ratio can easily be estimated from a simple retrieval of public

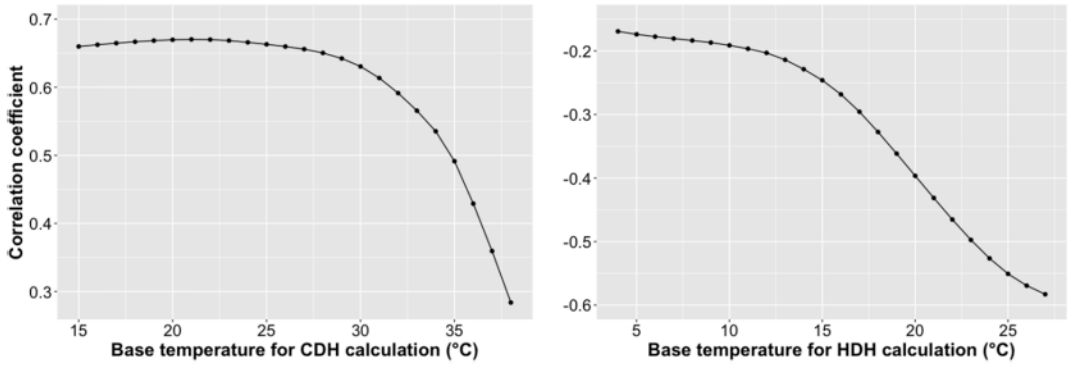

Fig. 4. Influence of base temperature on the correlation coefficient between degree-hours and space cooling energy usage per unit area.

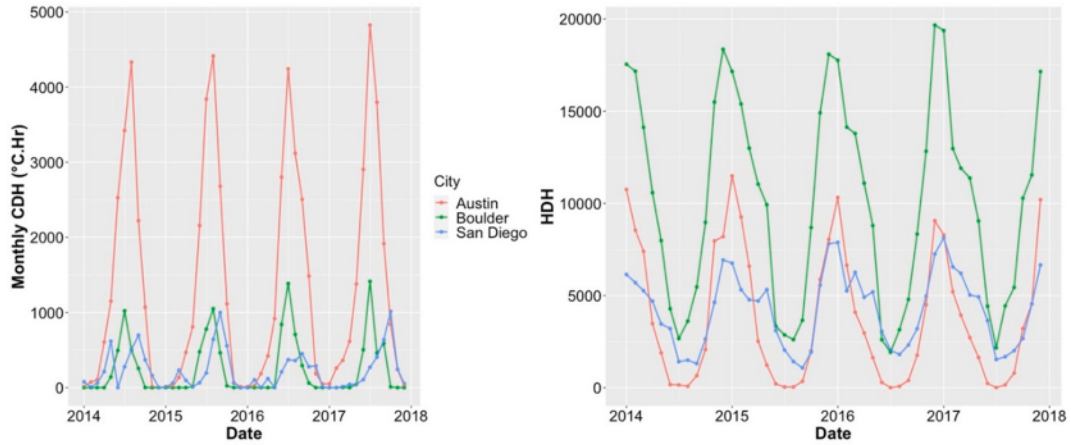

Fig. 5. Monthly accumulated CDH and HDH from 2014 to 2017.

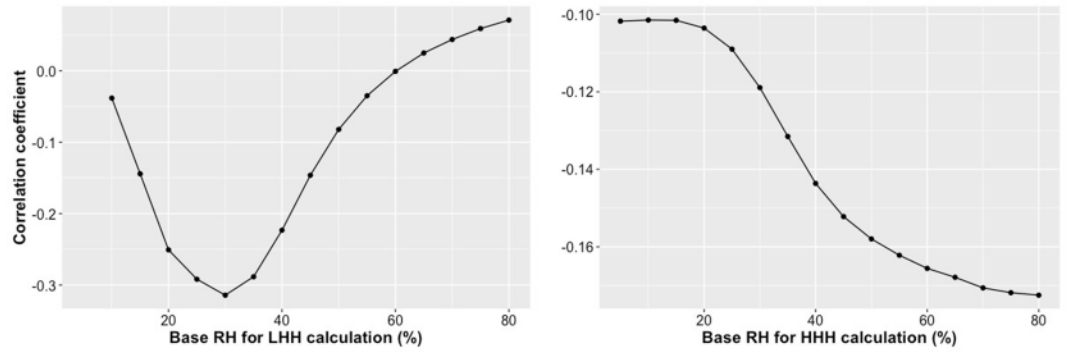

Fig. 6. Influence of base RH on the correlation coefficient between LHH and space cooling energy usage and HHH and space cooling energy usage.

property records that provide the floor area and the number of floors. The approximate calculation of S/V is expressed in Eq. (5). The floor-to-floor height of 3 m is assumed to be the same for each building; the floor shape was assumed roughly to be square. The values calculated for the S/V ratio for the sample buildings examined in this work are displayed in Fig. 8.

$$S/V = \frac{\sqrt{\frac{A_{floor}}{f}} \times 4 \times 10 + \frac{A_{floor}}{f}}{A_{floor} \times 10} \quad (5)$$

where A_{floor} is the total floor area and f is the number of floors for each building.

2.3. Preparation and initial analysis of candidate predictor variables

The input features include the variables obtained directly from the downloaded data and proposed variables that cover utility bills, building-specific data, and weather data. The initial input variables selected for cooling energy usage estimation are shown in Table 1. The monthly total energy usage per square foot can be obtained from utility bills. The building age and S/V ratio were used as representative characteristics of a building. The weather conditions in terms of temperature, humidity, solar radiation, and wind speed were fully considered. The parameters of temperature, as well as the proposed CDH and HDH, were selected as candidate parameters because they had different physical meanings, and the evaluation of their relative importance is one aim of this research. The same applies to the parameter of humidity, LHH, and HHH.

Pearson's correlation coefficients between the model variables were used to check the strength of their collinearity between predictors (See Fig. 9) firstly before developing models. This is to present the most basic

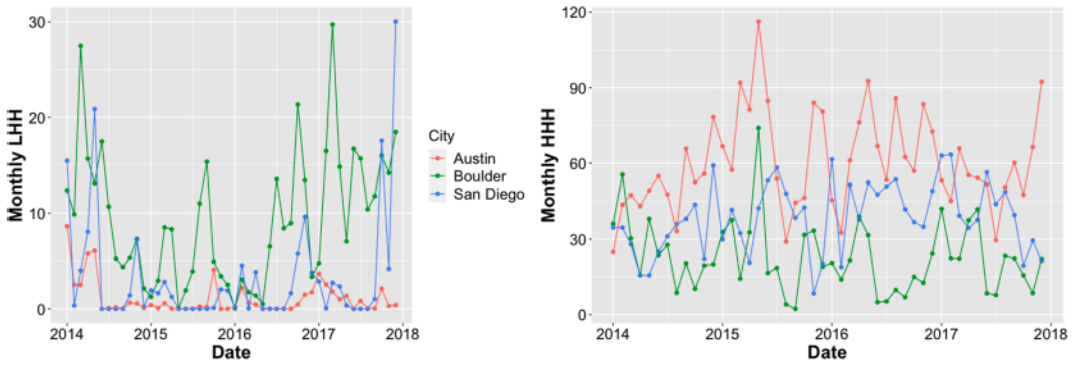

Fig. 7. Monthly accumulated LHH and HHH from 2014 to 2017.

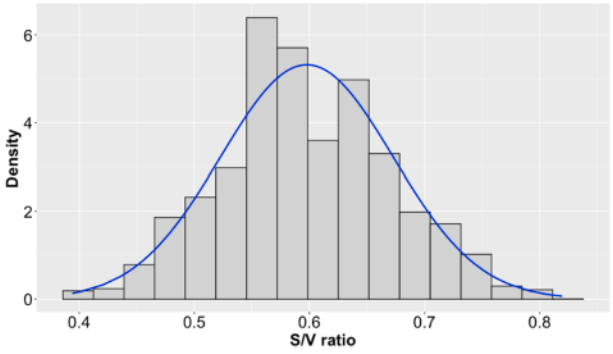

Fig. 8. Distribution of the S/V ratios for the sample buildings.

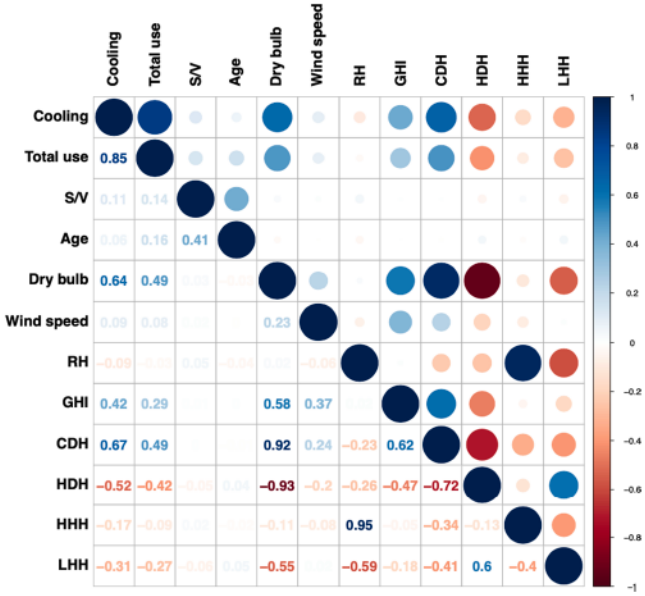

Fig. 9. Correlation matrix of predictors and responses for space cooling energy estimation.

Table 1
Symbol and definition of the input Parameters for Cooling Model Analysis.

Parameter	Symbol	Definition	Unit
Cooling	E_c	Monthly cooling energy use per unit area	kWh/sq·m
Total use	E_{tc}	Total monthly energy use per unit area	kWh/sq·m
Dry bulb	T_d	Monthly average dry bulb temperature	°C
CDH (T > 23.8)	h_{cd}	Cooling degree hours	°C·Hr
HDH (T < 23.8)	h_{hd}	Heating degree hours	°C·Hr
Relative humidity	RH	Monthly average relative humidity	%
HHH (RH > 70%)	h_{hh}	High humidity hours	Hrs
LHH (RH < 30%)	h_{lh}	Low humidity hours	Hrs
Solar radiation	GHI	Monthly global horizontal irradiance	Wh/m^2
Wind speed	W_s	Monthly average wind speed	m/s
S/V	S/V	Surface-area-to-volume ratio	m^{-1}
Age	Y_r	Age of house	Year

and background information and relationship among the parameters used in the next modeling processes. A positive correlation exists between two variables when the coefficient is positive, whereas a negative correlation indicates a negative coefficient value. It should be noted that a causal relationship is not implied. As the initial step, a correlation analysis enables to visually examine the possible relationships between the candidate predictors. It is important to note that the subsequently applied models also had built-in algorithms for ranking the variables and facilitating the parameter significance analysis for the candidate models.

2.4. Application of machine learning methods and performance evaluation metrics

Statistical regression methods and machine learning techniques were applied to develop the prediction models for space cooling energy usage per square foot. All of the predictors were numerical variables. The generalized linear model (GLM) and polynomial model were selected as representative regression models to express the prediction model using a simple interpretable equation. Machine learning techniques, including SVM, KNN, random forest, eXtreme gradient boosting (XGBoost), GAMBoost, and GLMBoost were also adopted in this study. These methods that are most commonly used in each algorithm balance their prediction biases and give full consideration to predictive accuracy. Earlier studies [21,23] examined the feasibility and applicability of SVM in forecasting the building energy consumption through investigating the influence of different SVM parameters on the estimation accuracy based on the SVM features such as mathematical kernel function defined and structure risk minimization regression used. These studies demonstrated the better prediction accuracy and model performance of SVM in building energy prediction than the traditional neural network model that had more free parameters to optimize. KNN method was implemented in [17] to capture the nonlinearities of the appliance-specific energy usage signals, and the study in [46] considered the inverse of Euclidean distance as the weight for the KNN model and proposed a short-term load prediction model with higher forecasting accuracy. The random forest method applies a tree-based algorithm to estimate energy consumption, and the study in [47] applied this homogeneous ensemble approach to predict hourly building electricity usage by examining the effect of parameter settings such as the numbers of variables on the

prediction accuracy. XGBoost is based on an optimized gradient boosting framework that can efficiently improve modeling accuracy [48]. Researchers in [49] used the XGBoost method on the time series data by feature selection to forecast the electricity load for a single time interval with efficient computing time while [50] proposed a hybrid model combining the XGBoost algorithm to evaluate the features' importance.

In this research, the data processing and machine learning analysis were conducted using the R package of "mlr3" available on CRAN [51]. This package is developed with a robust object-oriented framework for machine learning algorithms. The random search algorithm, which is an efficient and effective technique for hyper-parameter optimization, was applied for parameter tuning using k-fold cross-validation to improve the models' performance [52]. K-fold cross-validation with the value five assigned to k was adopted in this work to randomly generate the training and test sets and to estimate the tuning parameters for optimal model performance. The candidate models were developed using the first fold as the testing data and the remaining four folds as the training data. Using R^2 for the model with parameter λ, prediction accuracy was computed using the test data. This process was repeated five times, until each of the five subsets was taken as the training data exactly once. Cross-validation accuracy was evaluated by averaging the k estimates of R^2 from each iteration. The calculation of R^2 and MSE from each iteration as well as k-fold CV accuracy were based on Eqs. (6)–(8), respectively.

$$R_i^2(\lambda) = 1 - \frac{\sum_j (y_j - \widehat{y}_{j(\lambda)})^2}{\sum_j (y_j - \overline{y}_j)^2} \quad (6)$$

$$MSE_i(\lambda) = (y_j - \widehat{y}_i)^2 \quad (7)$$

$$CV(\lambda) = \frac{1}{k}\sum_{i=1}^{k} MSE_i(\lambda) \quad (8)$$

where y_j is the observed response value, $\widehat{y}_j(\lambda)$ is the predicted response value for the parameter λ, and $\overline{y}_j$ is the mean value of the responses.

The value of λ that made the CV (λ) value the largest was selected for optimal model development in each candidate model algorithm. The optimal models were then developed for the whole dataset. The optimal models' R^2 coefficients resulting from the k-fold cross-validation and using the whole dataset were compared to evaluate the models' validity. The prediction accuracy of all of the optimal energy usage models developed with the whole dataset was evaluated using the R^2, RMSE, and RAE. The final model was obtained from the candidate optimal models after a comprehensive comparison. The feature selection and importance of the final model were identified and compared to determine the most influential factors in each model for space cooling energy use. The model development framework is summarized in Fig. 10.

3. Results

3.1. Model development and performance comparison

The polynomial model, GLM, and machine learning models were then developed. Table 2 summarizes the models' prediction performances according to the criteria of R^2, MSE, RMSE, and RAE. The models' R^2 coefficients and MSE were obtained via the cross-validation approach, which were named CV_R^2 and CV_MSE, respectively. The optimal models built from the whole dataset were selected mainly based on MSE. To more clearly understand and evaluate the best model, the bar plot in Fig. 11 was developed to depict the models' performances.

Table 2
Prediction Performances of the Nine Models Developed using various evaluation metrics.

Method	CV_R2	CV_MSE	M_R2	MSE	RMSE	RAE
polynomial	0.818	0.519	0.820	0.514	0.717	0.382
glm	0.802	0.566	0.802	0.565	0.752	0.403
xgboost	0.922	0.221	0.970	0.086	0.294	0.153
ranger	0.852	0.420	0.977	0.065	0.256	0.132
cforest	0.829	0.489	0.862	0.395	0.629	0.331
knn	0.806	0.553	0.869	0.373	0.611	0.332
svm	0.836	0.470	0.855	0.413	0.643	0.336
gamboost	0.833	0.478	0.848	0.435	0.659	0.351
glmboost	0.804	0.562	0.805	0.557	0.746	0.399

The XGBoost model had the second-lowest MSE value with 0.086 and a maximum value of CV_R^2, approximately 92.2%. The small difference between CV_MSE and MSE with the value of 0.135, and between the values of CV_R^2 and M_R^2 indicated a low risk of overfitting. The low RMSE and RAE values also indicated good predictive accuracy. A ranger model was obtained with the lowest MSE value of 0.065 and a maximum R^2 value of around 97.7%; however, that model had a much larger CV_MSE value and lower CV_R^2 value of 85.2%, showing that much uncertainty or bias could be introduced into the prediction by using the whole dataset, reducing the model's reliability and validity. Average predictive performance was found in the SVM, cforest, and GAMBoost models, and less predictive accuracy was obtained from the GLM, glmboost, and polynomial models.

The model performance also differs in terms of computational resources, including computation time and memory space usage. The generated polynomial model and GLM models took much less time than that it took to attain the machine learning models, but these two models had a relatively lower prediction performance with R^2 around 0.8. The machine learning model development relied much on the hardware for a smooth and quick computation, and the average time for developing the XGBoost model in this study was around 30 min. On the other hand, there was not a significant difference in the computation time among these machine learning models when using the package "mlr3". Comparatively speaking, the XGBoost model with much better prediction performance is worthwhile at the cost of acceptable computing time and efforts.

Through comparison and analysis of these models, the XGBoost model was finally selected as the optimal model for this modeling scenario. The XGBoost model implements parallel tree learning and allows users to custom evaluation criteria, and also it has a built-in cross-validation procedure at each boosting iteration [48]. The tree-specific parameters that were optimized using cross-validation for overfitting and under-fitting control in this study were shown in Table 3.

3.2. Application of the XGBoost model and the prediction performance

The XGBoost model was applied to the collection of datasets, of which the data for Austin comprised the most substantial proportion.

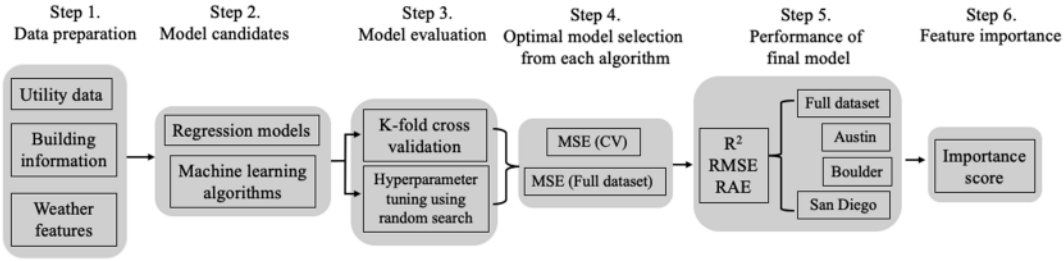

Fig. 10. Model development framework.

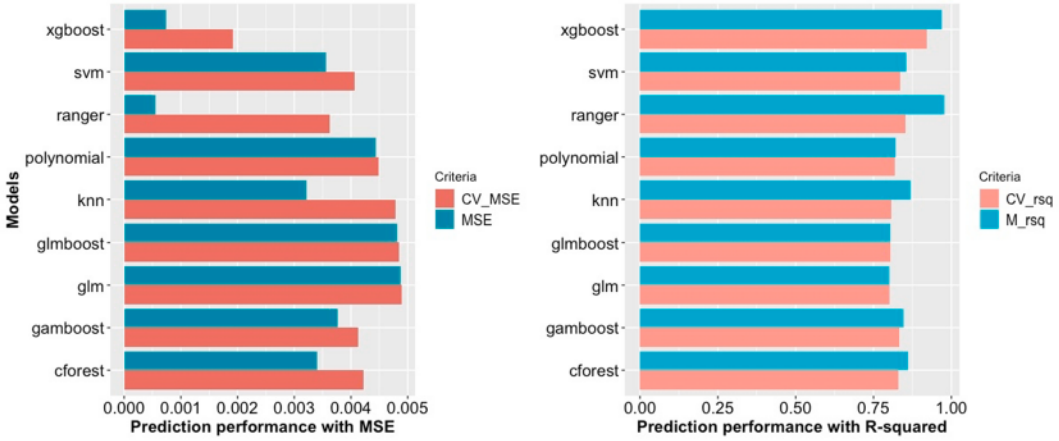

Fig. 11. Prediction performance comparison plots for selected algorithms.

Table 3
Tree-specific parameters tuning for XGBoost model.

Tree-specific parameters	max_depth	gamma	min_child_weight	subsample	colsample_bytree
Value	5	0	1	0.9	0.9

The regression line of the truth and response values are shown in Fig. 12a. The whole dataset included data for three cities, and thus the model's applicability for each individual city was examined by relationship plots (see Fig. 12b, 12c, and 12d for Austin, Boulder, and San Diego, respectively). The diagonal lines' adjusted R^2 values indicated the goodness-of-fit of the XGBoost model for the three cities in different climate zones.

The model was then evaluated in terms of its prediction accuracy for each individual building (defined by a specific building ID). There were 377 houses with multiple observations, 324 of them with a minimum of 10 observations. These were sorted out as representative buildings for the predictive analysis, buildings with too few observations may have yielded unwanted errors. Fig. 13 shows the distribution of the R^2 values for each building. The model appeared unsuitable for three buildings with negative R^2 values (not shown in Fig. 13). Other than those negative values, R^2 values less than 80% for 18 buildings were obtained. The prediction accuracy for 13 buildings ranged between 80% and 85%. That is to say that using the model developed for this research, 290 buildings, 90% of the dataset, could be evaluated with no less than 85% prediction accuracy in terms of their cooling energy usage.

3.3. Feature importance in the XGBoost model

The model obtained consisted of the parameters of utility bills, building-specific information, and weather-related data. The relative importance level of each input variable was evaluated to determine the major factors influencing cooling energy estimation. The input variables were compared and ranked using feature importance scores intelligently obtained by the boosted trees of the XGBoost algorithm [48,51]. Feature importance is shown in Fig. 14a, which indicates that the total energy usage obtained from utility data was the most influential predictor variable for cooling energy estimation. The variable of secondary importance was the S/V ratio, which represents the critical building characteristic of compactness. Almost equal significance was found for

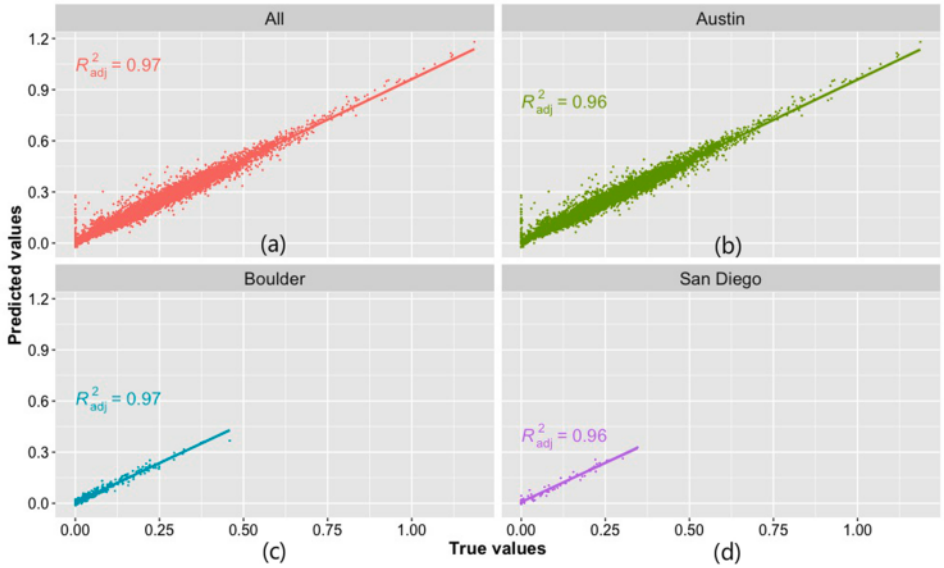

Fig. 12. True values vs. predicted response values.

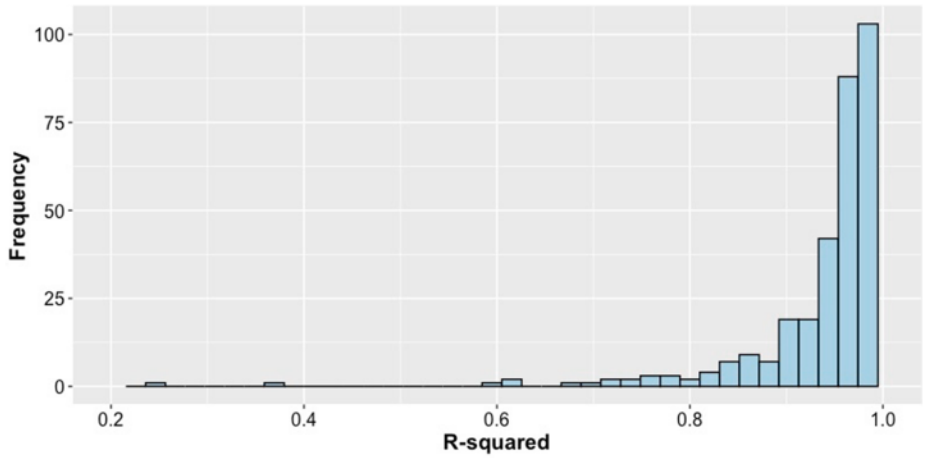

Fig. 13. Distribution of prediction accuracy for each building.

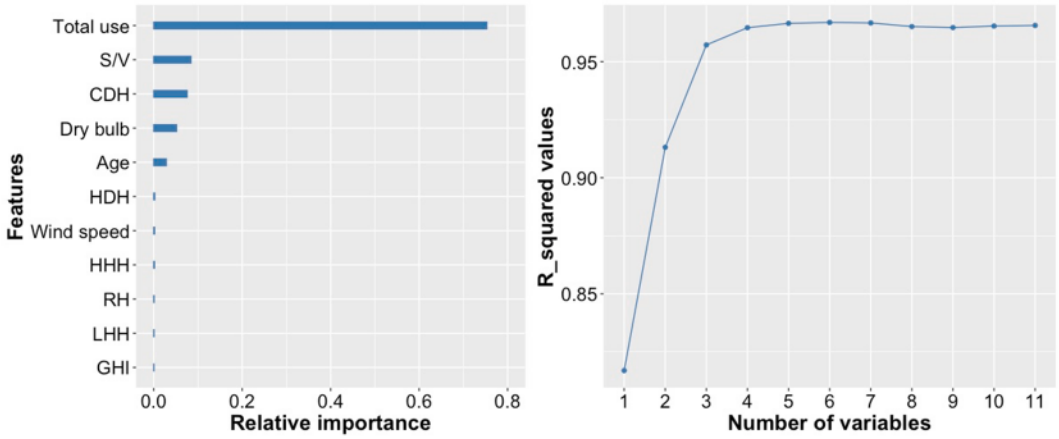

Fig. 14. (a) Feature importance, and (b) the influence of the number of selected variables on the R^2.

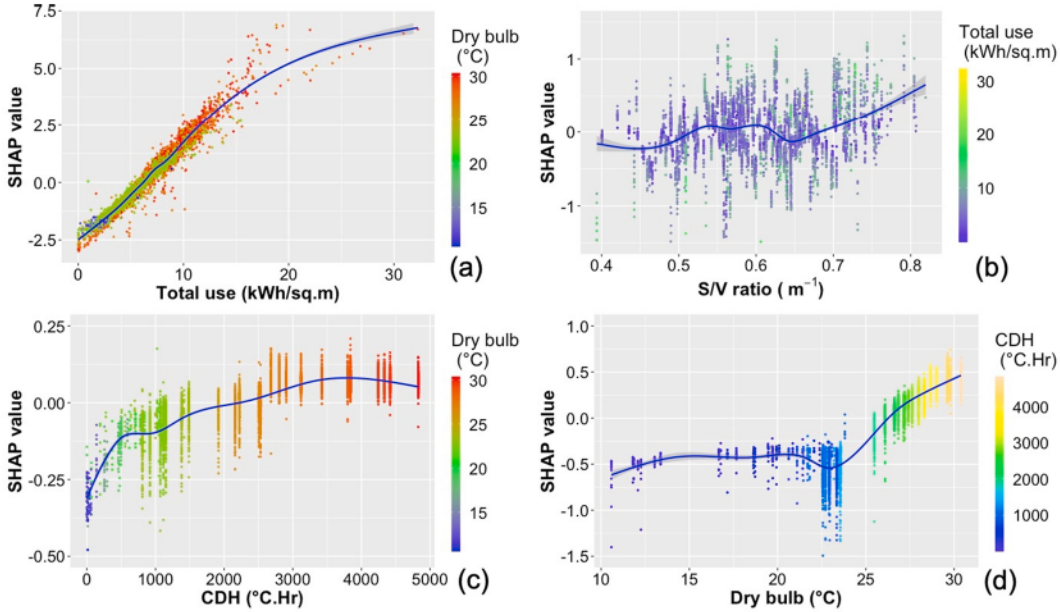

Fig. 15. SHAP dependence plot for significant features.

the CDH variable, which outperformed DBT. Thus, CDH would well reflect temperature information and its influence on a building's cooling energy demand. Building age was found to be another relatively important parameter influencing building cooling energy consumption. Other weather-related variables, including RH, humidity-related degree hours, wind speed, and GHI were of little importance compared to the other variables analyzed above. Further analysis was conducted to consider the influence of the number of input variables in the model on its predictive accuracy (see Fig. 14b). The variables selected for the model with different numbers of input predictors were based on their relative importance scores based on gradient boosting algorithm [48,53]. The maximum R^2 at approximately 97% could be reached with a minimum of four input parameters: total energy use, S/V ratio, CDH, and DBT.

The developed model has accurate prediction accuracy, but it is still necessary to understand and interpret the relationship between each of the four features and cooling energy usage in this model. The interpretability of complex models would lead to a better understanding of how the model makes a decision and thus enhance model acceptance and adoption. The SHapley Additive exPlanations (SHAP) approach, which is deemed as an advanced method to interpret individual prediction for tree-based models, was employed to display the relationship by computing the contribution of each variable to the expected model prediction [54]. Fig. 15 shows the SHAP dependence plot to illustrate the effects of the key features on the response prediction interacted with other features. The horizontal axis represents the actual value of the feature being depicted. The vertical axis, the SHAP value, represents how much of the contribution of that feature's value to the model prediction output. The SHAP dependence also shows the variance on the y-axis. Especially in the case of interactions, the SHAP dependence plot will be much more dispersed in the y-axis. The color corresponds to a second feature that may have an interaction effect with the feature we selected. As shown in Fig. 15a, the overall energy usage's SHAP value distributions indicate its strong effects on the model, which aligns with the results of feature relative importance analysis above. The upward trend of the slopes indicates a positive and nonlinear relationship between the overall energy use and the target variable – cooling energy use, which is also consistent with the basic correlation analysis in section 2.3. Also, the cluster of greenish colored dots in the range of 0–10.8 total energy use shows a possible interaction effect of the total energy use with the dry bulb, which will be further analyzed in the next part. The SHAP dependence plot for the S/V ratio in Fig. 15b has a large dispersion in the vertical direction, especially in the low S/V region (<0.66), which suggests that this building shape-related feature may affect the prediction by interacting considerably with the other variables. The relationship between the cooling energy use and S/V in this region is monotonic. However, intuitively, we know under the same volume conditions, the greater the building surface area, the greater the potential heat gain or loss through it. This relationship is well captured in the high S/V region (>0.66) as there is a relatively important and positive correlation between the S/V and cooling energy use in this region. Similarly, in the plot of Fig. 15c, there exists a positive nonlinear relationship between CDH and cooling energy use. Meanwhile, the variation of dry bulb temperature was colored, and the color pattern aligns with the definition of CDH in which the higher the dry bulb temperature, the larger the CDH value will be. Dry bulb temperature and its SHAP values in Fig. 15d show an interesting pattern in which the dry bulb temperature seems to be an important feature that negatively affects the model prediction value when it is below approximately 23.8 °C (75°F), while the relationship between dry bulb temperature and its effects on cooling energy use is monotonic. However, there exists a linear and positive relationship between the feature and the target when the dry bulb temperature is higher than 23.8 °C (75°F). This also echoes the earlier finding on the use of 23.8 °C (75°F) as the CDH base temperature.

As mentioned above, we noticed that potential interaction exists between features – the dry bulb temperature and the total energy use at the energy use region (0–10.8). Therefore, after accounting for the individual feature effects, we also studied the interaction effect (or called the combined feature effect) on the model output by using the SHAP interaction values. Such values are computed after subtracting the main effect of the features so that the interaction effect between features on the model can be obtained [55]. Fig. 16 presents the SHAP interaction value plots for the feature pair of the total energy use and the dry bulb temperature. By plotting the SHAP interaction values, we can visualize that the total use interacted with the dry bulb temperature range 18.0–23.8 °C (65–75°F) (greenish dots) has a rapidly sloping distribution (negative relationship to the prediction) and relatively higher SHAP values mainly stay in the 0–10 total energy use range. The same interaction effect is not visualized for any other temperature regions. This finding using the SHAP interaction value framework is fully comprehensible because the increase of the total energy use should not be mostly contributed by the cooling use when the outdoor temperature is within the comfort zone (18.0–23.8 °C) (65–75°F). Also, this information is supportive of our earlier pre-processing of the cooling-dominant data.

4. Discussion

The typical machine learning algorithms were adopted and compared in this study. Hyperparameter tuning was applied to each method for exploiting their optimal performance and making them comparable with each other. Also, k-fold cross-validation was utilized in the process of solving the hyperparameters to reduce the risk of overfitting and increase the model's validity. Through the model development and comparison, we identified problems and limitations related to the input variables that affect the accuracy of the prediction, as well as opportunities to apply this model to solve real-world building energy performance issues.

4.1. The reliability and consistency of building energy usage data

It is important to note that the sample buildings selected in this study were located in three different climate zones (i.e., Zones 2A, 3B, and 5B), which was an intentional effort to take various weather-related parameters into account. However, the number of observations for the sample buildings located in the three target cities were not equal due to limitations on the quantity of qualified data. This may have had an adverse effect on the predictive performance of buildings located in various climates. The results of applying the XGBoost model on individual cities showed a 1% lower prediction accuracy for the city of San Diego. Further research should be conducted to increase the observations and balance the number of sample buildings from different cities to improve prediction accuracy.

The data pre-processing, including data cleaning, data transformation, and the feature candidate selection was firstly performed before model development. The process of data cleaning was intended to remove unqualified data such as extremely low cooling energy consumption data in cold seasons to ensure the quality of the collected data. For example, the results of the prediction accuracy of the XGBoost model for each individual building showed negative R^2 values for three buildings. These three buildings were found to have far lower monthly cooling energy usage compared to buildings with similar building characteristics. The unusual cooling energy consumption indicates that there may be uncontrolled or unconsidered factors, such as individual living habits and occupancy information, that affect the cooling needs. For the 18 buildings with R^2 values less than 80%, it was found that there were large variations of monthly cooling energy usage for a building with similar monthly total energy usage or a significant difference in monthly total energy usage with almost the same cooling energy usage. The data consistency for these 18 buildings' energy usage needs further evaluation. It is an important work to identify the cooling-dominant data and remove unqualified data in the data pre-processing

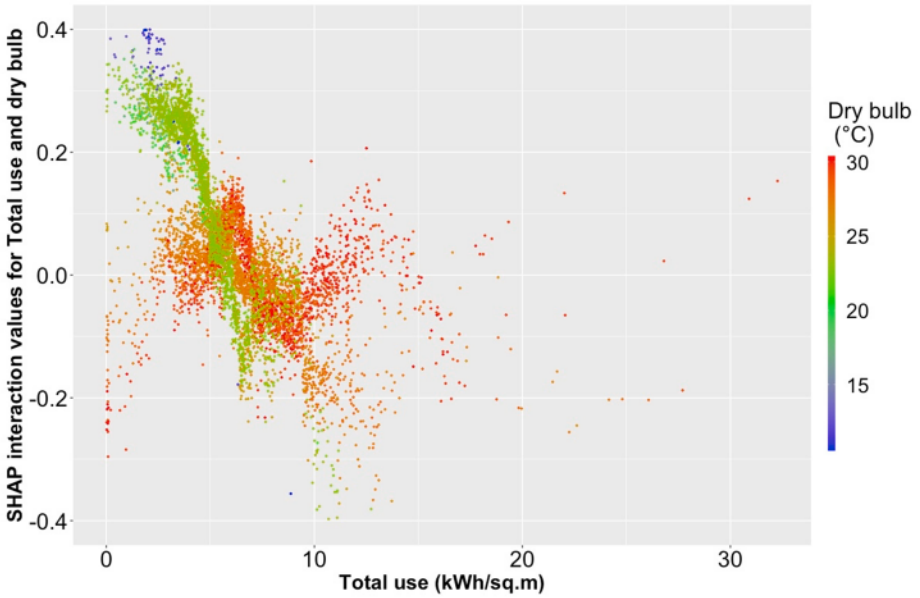

Fig. 16. SHAP interaction values for interpreting the interaction effect.

phase to ensure data reliability and a more effective modeling procedure. The problem of model application in cold seasons would be possibly overcome by integrating more reliable measured data in low-temperature conditions for a building to enhance model generalization or by incorporating a decision-tree-based structure to determine whether the cooling energy use occurs before applying this cooling use estimation model. Overall speaking, this developed prediction model could support occupants understand the influence of their behaviors on building energy performance and thus promote energy-saving behaviors.

4.2. Evaluation and comparison of parameters' importance in various models

These different modeling methods demonstrated consistency in terms of feature importance. In the XGBoost model, cooling energy use was found to be more sensitive to the CDH parameter than to DBT. Comparatively, the RH parameter and proposed humidity hours played insignificant roles in this model. The GHI variable was also not particularly sensitive to cooling energy prediction, which seems inconsistent with the first principles of building physics. To further explore the possibility of taking GHI into account, we roughly estimated the solar radiation received by the building envelope by using the envelope's area and GHI. Unfortunately, this did not improve the correlation coefficient. The reason might be that the detailed building characteristics information related to the solar radiation received (such as building orientation, window-to-wall ratio, and tree shading) dramatically affects the actual solar heat gain received but remains obscure in our modeling process. In other words, more information is needed to identify the influence of solar radiation. However, it can be envisioned that successfully adding solar radiation to the model would be a key factor in boosting its performance.

In addition to these general reflections, we chose the four best models to further analyze with regards to feature importance, including XGBoost, Ranger, GAMBoost, and cforest. Fig. 17 displays the relative importance of the most influential variables in these four models. The importance scores for each model were normalized, and the relative importance of the features was compared. The top six ranked variables

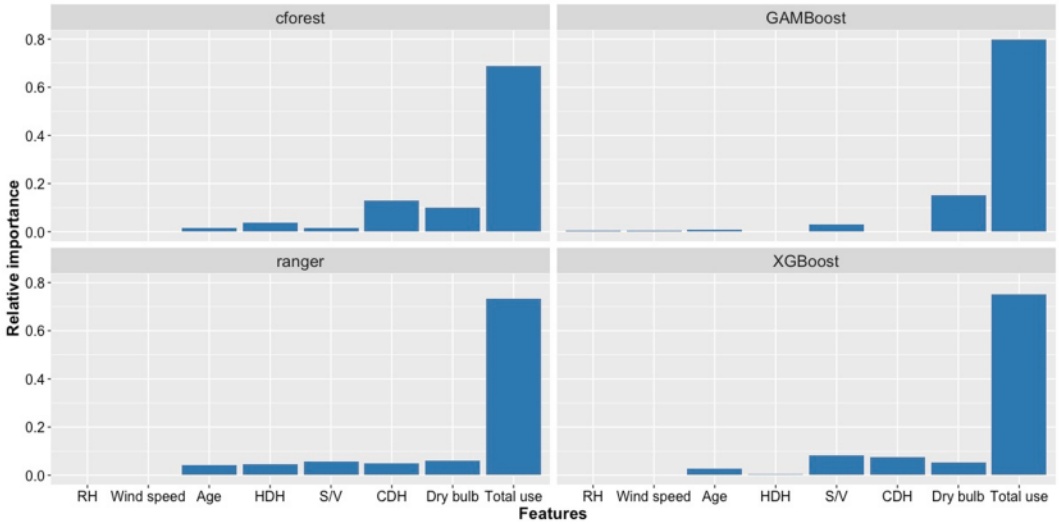

Fig. 17. Relative importance of features in the comparison models.

for each model were selected, and it was discovered that the importance ranking of the variables differed by modeling technique. However, the variable of total use was always the most important predictor. DBT and CDH remained the second most important in the cforest, ranger, and XGBoost models. As discussed previously, based on building physics knowledge, the building S/V ratio should have a significant effect on cooling energy consumption; a compact building shape will benefit energy savings. Interestingly, the significance of the S/V ratio was only detected in the ranger and XGBoost models, which might be the reason behind their improved accuracy over all other models. Overall, the model performance would benefit from the integration of the processed CDH and S/V variables when predicting cooling energy usage (in addition to the weather features traditionally employed).

4.3. Workflow implementation

The model identified and developed here can be written via a programming language to form an executable file, which could then be used by homeowners seeking to analyze their energy use. Fig. 18 depicts a schematic for using this model. Homeowners need only to input a selected month's overall home energy use data from their utility bills and their physical home address. Based on the geographic location, two data extraction and computation processes would be carried out. First, the historical weather data in that particular month (consistent with the utility bill's month) would be retrieved from public databases such as the National Weather Service and National Oceanic and Atmospheric Administration. Then, the CDH parameter would be calculated based on Eq. (1) and an assumed base temperature of 23.8 °C (75°F). Second, the physical address input by the homeowner would be used to retrieve the building characteristics from public record databases of real property (handled by individual counties or cities, or property search engines such as US Reality Records). The S/V and age parameters would then be output according to the building profile-related data extracted, including the number of floors, overall area, and year built. Overall energy use, weather-related features, and building characteristics would then be used to yield the cooling energy use in the month indicated.

Building upon the aforementioned information, various external calculations, and strategies related to energy savings could be combined and used for a wide variety of purposes. For instance, in our previous work, we developed an *in situ* window measurement sensor module that can conveniently be assembled and used to report key home window properties, including the thermal coefficient, visible transmittance, and solar transmittance [56]. Combining information related to window properties and space cooling energy use in a particular month, we should be able to provide more accurate information regarding energy use and cost percentages home windows may account for, allowing for the presentation of comparable quantities obtained by upgrading the windows or window films. Consider, for example, behavioral intervention strategies designed and implemented to increase household energy-saving behaviors by providing normative energy use feedback [57,58]. More detailed energy disaggregation data (such as cooling energy use data) would significantly enhance the effectiveness of these strategies by identifying more meaningful reference groups and providing more personalized feedback.

5. Conclusions

This study compared several statistical modeling techniques and developed a reliable XGBoost model that accurately predicts cooling energy use for residential buildings using the input variables of utility bill information, building surface-to-volume ratio, cooling degree hours, and dry-bulb temperature. The k-fold cross-validation and parameter tuning were applied to evaluate the model's reliability and validity. The XGBoost model developed offers optimal performance compared to the other commonly used statistical methods. An R^2 value of 92% was obtained to estimate the accuracy using k-fold cross-validation. This model was also individually applied to cities located in three different climate zones. The results indicated good reliability of prediction and goodness-of-fit.

Furthermore, several new weather- and building characteristics-related parameters were designed and tested in the modeling procedure, among which predictors of cooling degree hours and surface-to-volume ratio were found to be effective in enhancing the model's prediction performance. Models using degree hours and surface-to-volume ratio exhibit a great potential to increase model performance for future energy prediction studies. Although some other newly designed parameters impacted the model results when specific modeling methods were used, they were not statistically significant for determining the space cooling energy use results in our research. It is possible that they may have a more beneficial effect in other climatic zones, such as with high humidity hours in extremely humid climates. This issue also informs one of this work's limitations.

To summarize, this work demonstrated the feasibility of excellent prediction of space cooling energy use based on readily available datasets and simple user input, without computation complexity or the need for additional hardware. To enhance the generalizability of the models, we intentionally incorporated local weather data and physical characteristics of the buildings into the modeling procedure; we also included three climatic zones. However, the model's performance may still be limited in terms of building and weather features in the database selected. More external validation is required to ensure the generalizability of the model. Additionally, the original datasets used to build this model stem from disaggregated hourly energy use datasets in the single-family house sector in which the implementation of the model is limited to. That being said, to expand the use of this model or follow modeling methods and predictors in other studies such as other residential typologies (e.g., apartments, dormitories, townhouses), similar levels of detail for the training dataset are still necessary.

CRediT authorship contribution statement

Yanxiao Feng: Data curation, Investigation, Formal analysis, Validation, Visualization. **Qiuhua Duan:** Data curation. **Xi Chen:** Resources. **Sai Santosh Yakkali:** Resources. **Julian Wang:**

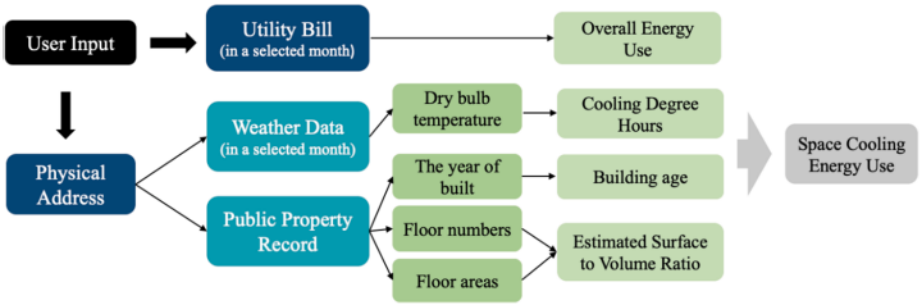

Fig. 18. Workflow of implementation.

Conceptualization, Methodology, Project administration.

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.

Acknowledgements

We acknowledge the financial support provided by the National Science Foundation CMMI- 2001207 and Environmental Protection Agency SU836940.

References

[1] U.S. energy information administration, residential energy consumption survey; 2015. Available: https://www.eia.gov/consumption/residential/index.php.
[2] U.S. Energy Information Administration, Annual Energy Outlook 2019; January 2019.
[3] Allcott H, Mullainathan S. Behavior and Energy policy. Science 2010;327(5970): 1204–5. https://doi.org/10.1126/science.1180775.
[4] Abrahamse W, Steg L, Vlek C, Rothengatter T. A review of intervention studies aimed at household energy conservation. J Environ Psychol 2005;25(3):273–91. https://doi.org/10.1016/j.jenvp.2005.08.002.
[5] Darby S. The effectiveness of feedback on energy consumption. Environmental Change Institute, University of Oxford; April 2006. Available: http://copperalliance.org.uk/uploads/2018/02/smart-metering-report.pdf.
[6] Neenan B, Robinson J. Residential electricity use feedback: A research synthesis and economic framework. Technical report. Electric Power Research Institute; 2009.
[7] Hart GW. Nonintrusive appliance load monitoring. Proc IEEE 1992;80(12): 1870–91. https://doi.org/10.1109/5.192069.
[8] Zeifman M, Roth K. Nonintrusive appliance load monitoring: Review and outlook. IEEE Trans Consum Electron 2011;57(1):76–84. https://doi.org/10.1109/TCE.2011.5735484.
[9] Zoha A, Gluhak A, Imran MA, Rajasegarar S. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. Sensors 2012;12(12): 16838–66. https://doi.org/10.3390/s121216838.
[10] Sultanem F. Using appliance signatures for monitoring residential loads at meter panel level. IEEE Trans Power Delivery 1991;6(4):1380–5. https://doi.org/10.1109/61.97667.
[11] Norford LK, Mabey N. Non-intrusive electric load monitoring in commercial buildings. Proceedings of the eighth symposium on improving building systems in hot and humid climates, Dallas, TX, USA. 1992.
[12] Laughman C, Lee K, Cox R, Shaw S, Leeb S, Norford L, et al. Power signature analysis. IEEE Power Energ Mag 2003;1(2):56–63. https://doi.org/10.1109/MPAE.2003.1192027.
[13] Perez KX, Cole WJ, Rhodes JD, Ondeck AD, Webber ME, Baldea M, et al. Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data. Energy Build 2014;81:316–25. https://doi.org/10.1016/j.enbuild.2014.06.031.
[14] Zia T, Bruckner D, Zaidi A. Hidden Markov model based procedure for identifying household electric loads. In: Proceedings of the 37th annual conference on IEEE industrial electronics society, Melbourne, Australia; 2011. p. 3218–322, doi: 10.1109/IECON.2011.6119826.
[15] Parson O, Ghosh S, Weal M, Rogers A. Non-intrusive load monitoring using prior models of general appliance types. Association for the Advancement of Artificial Intelligence (AAAI); 2012.
[16] Çavdar İH, Faryad V. New Design of a supervised energy disaggregation model based on the deep neural network for a smart grid. Energies 2019;12(7). https://doi.org/10.3390/en12071217.
[17] Khan MM, Siddique MA, Sakib S. Non-intrusive electrical appliances monitoring and classification using K-nearest neighbors; 2019. Available: arXiv:1911.13257.
[18] Kolter JZ, Johnson MJ. REDD: A public data set for energy disaggregation research; 2011. Available: http://redd.csail.mit.edu/kolter-kddsust11.pdf.
[19] Sonderegger RC. A baseline model for utility bill analysis using both weather and non-weather-related variables. ASHRAE Trans 1998;104.
[20] Dhar A, Reddy TA, Claridge DE. A fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. Trans ASME J Sol Energy Eng 1999;121(1):47–53. https://doi.org/10.1115/1.2888142.
[21] Dong B, Cao C, Siew EL. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37(5):545–53. https://doi.org/10.1016/j.enbuild.2004.09.009.
[22] Cho SH, Kim WT, Tae CS, Zaheemddin M. Effect of length of measurement period on accuracy of predicted annual heating energy consumption of buildings. Energy Convers Manage 2004;45(18–19):2867–78. https://doi.org/10.1016/j.enconman.2003.12.017.
[23] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Applying support vector machine to predict hourly cooling load in the building. Appl Energy 2009;86(10):2249–56. https://doi.org/10.1016/j.apenergy.2008.11.035.
[24] Rai P, Nassif N, Eaton K, Rodrigues A. Applications of machine learning in building energy prediction and savings. Energy Res J 2019;10(1). https://doi.org/10.3844/erjsp.2019.1.10.
[25] Zeng A, Ho H, Yu Y. Prediction of building electricity usage using Gaussian Process Regression. J Build Eng 2020;28. https://doi.org/10.1016/j.jobe.2019.101054.
[26] Better buildings accelerators. U.S. Department of Energy, Energy data access: Blueprint for action toolkit. https://betterbuildingssolutioncenter.energy.gov/toolkits/energy-data-access-blueprint-action-toolkit.
[27] Better buildings of U.S. Department of Energy, Home Energy Score Scoring Methodology; Feb 2017. Available: https://betterbuildingssolutioncenter.energy.gov/sites/default/files/attachments/Home%20Energy%20Score%20Methodology%20Paper%20v2017.pdf.
[28] Yan CC, Wang SW. A simplified energy performance assessment method for existing buildings based on energy bill disaggregation. Energy Build 2012;55: 563–74. https://doi.org/10.1016/j.enbuild.2012.09.043.
[29] Park JS, Lee SJ, Kim KH, Kwon KW, Jeong JW. Estimating thermal performance and energy saving potential of residential buildings using utility bills. Energy Build 2016;110:23–30. https://doi.org/10.1016/j.enbuild.2015.10.038.
[30] Dong B, Lee SE, Sapar MH. A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore. Energy Build 2005; 37(2):167–74. https://doi.org/10.1016/j.enbuild.2004.06.011.
[31] Pecan Street Dataport. Available: https://www.pecanstreet.org/dataport/.
[32] Andreas A, Stoffel T. NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS). Golden, Colorado (Data). NREL Report No. DA-5500-56488; 1981. doi: 10.5439/1052221.
[33] Badescu V, Zamfir E. Degree-days, degree-hours and ambient temperature bin data from monthly average temperatures (Romania). Energy Convers Manage 1999;40 (8):885–900. https://doi.org/10.1016/S0196-8904(98)00148-4.
[34] Papakostas K, Kyriakis N. Heating and cooling degree-hours for Athens and Thessaloniki. Greece, Renewable Energy 2005;30(12):1873–80. https://doi.org/10.1016/j.renene.2004.12.002.
[35] Bolattürk A. Optimum insulation thicknesses for building walls with respect to cooling and heating degree-hours in the warmest zone of Turkey. Build Environ 2008;43(6):1055–64. https://doi.org/10.1016/j.buildenv.2007.02.014.
[36] Yakkali SS, Feng Y, Chen X, Chen Z, Wang J. Estimating home heating and cooling energy use from monthly utility data. In: Proceedings of the future technologies conference. Cham: Springer; 2020. p. 124–34.
[37] ASHRAE. Handbook 2017: Fundamentals, SI edition. Atlanta (GA, USA): American Society of Heating, Refrigeration and Air Conditioning Engineers, Inc.; 2017.
[38] Shi Y, Han ZY, Xu Y, Xiao C. Impacts of climate change on heating and cooling degree-hours over China. Int J Climatol 2020. https://doi.org/10.1002/joc.6889.
[39] Winkler J, Munk J, Woods J. Sensitivity of occupant comfort models to humidity and their effect on cooling energy use. Build Environ 2019;162. https://doi.org/10.1016/j.buildenv.2019.106240.
[40] Zevenhoven R, Erlund R, Tveit T-M. Energy efficiency of exhaust air heat recovery while controlling building air humidity: A case study. Energy Convers Manage 2019;195. https://doi.org/10.1016/j.enconman.2019.05.058.
[41] Araji MT. Surface-to-volume ratio: How building geometry impacts solar energy production and heat gain through envelopes. IOP Conf Ser: Earth Environ Sci 2019; 323.
[42] Al-Saggaf A, Nasir H, Taha M. Quantitative approach for evaluating the building design features impact on cooling energy consumption in hot climates. Energy Build 2020;211. https://doi.org/10.1016/j.enbuild.2020.109802.
[43] Chen KW, Janssen P, Schlueter A. Multi-objective optimisation of building form, envelope and cooling system for improved building energy performance. Autom Constr 2018;94. https://doi.org/10.1016/j.autcon.2018.07.002.
[44] Wang JJ, Beltran L. A method of energy simulation for dynamic building envelopes. Proc SimBuild 2016;6(1).
[45] Li J, Duan Q, Zhang E, Wang J. Applications of shape memory polymers in kinetic buildings. Adv Mater Sci Eng 2018. https://doi.org/10.1155/2018/7453698.
[46] Fan GF, Guo YH, Zheng JM, Hong WC. Application of the weighted K-nearest neighbor algorithm for short-term load forecasting. Energies 2019;12(5). https://doi.org/10.3390/en12050916.
[47] 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.
[48] Tian C, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16). Association for Computing Machinery, New York, NY, USA; August 2016, doi: 10.1145/2939672.2939785.
[49] Abbasi RA, Javaid N, Ghuman MNJ, Khan ZA, Ur Rehman S, Amanullah. Short term load forecasting using XGBoost, In: Web, artificial intelligence and network applications. WAINA 2019. Advances in intelligent systems and computing, vol. 927. Cham: Springer; 2019, doi: 10.1007/978-3-030-15035-8_108.
[50] Zheng HT, Yuan JB, Chen L. Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. Energies 2017;10(8). https://doi.org/10.3390/en10081168.
[51] Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, et al. Bischl, mlr3: A modern object-oriented machine laearning framework in R. J Open Source Software 2019;4(44). https://doi.org/10.21105/joss.01903.
[52] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305.

[53] Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232.
[54] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017.
[55] Lundberg SM, Erion CG, Lee SI. Consistent individualized feature attribution for tree ensembles; 2019. arXiv: 1802.03888.
[56] Feng YX, Duan QH, Wang J, Baur S. Approximation of building window properties using in situ measurements. Build Environ 2020;169. https://doi.org/10.1016/j.buildenv.2019.106590.
[57] Khosrowpour A, Jain RK, Taylor JE, Peschiera G, Chen JY, Gulbinas R. A review of occupant energy feedback research: Opportunities for methodological fusion at the intersection of experimentation, analytics, surveys and simulation. Appl Energy 2018;218:304–16. https://doi.org/10.1016/j.apenergy.2018.02.148.
[58] Anderson K, Song K, Lee SH, Krupka E, Lee H, Park M. Longitudinal analysis of normative energy use feedback on dormitory occupants. Appl Energy 2017;189:623–39. https://doi.org/10.1016/j.apenergy.2016.12.086.