University of Reading

Modelling heating and cooling energy demand for building stock using a hybrid approach

Article

Accepted Version

Li, X. and Yao, R. ORCID: https://orcid.org/0000-0003-4269-7224 (2021) Modelling heating and cooling energy demand for building stock using a hybrid approach. Energy and Buildings, 235. 110740. ISSN 0378-7788 doi: https://doi.org/10.1016/j.enbuild.2021.110740 Available at https://centaur.reading.ac.uk/98368/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: http://dx.doi.org/10.1016/j.enbuild.2021.110740
To link to this article DOI: http://dx.doi.org/10.1016/j.enbuild.2021.110740

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading
Reading's research outputs online

Modelling heating and cooling energy demand for building stock using a hybrid approach

Xinyi Li [1,2*], Runming Yao[2,3]

[1,] Department of Civil and Structural Engineering, University of Sheffield, Sheffield, S1 3JD, UK;

[2,] Joint International Laboratory of Green Buildings and Built Environments, Ministry of Education, Chongqing University, Chongqing, 400045, China;

[3,] School of the Built Environment, University of Reading, Whiteknights, Reading, Whiteknights PO Box 219, RG6 6AW, UK;

Highlights

- A hybrid approach for building stock energy prediction
- An energy prediction model for both residential and non-residential buildings
- Prediction performance comparison of ten machine learning models
- The best performed model at building and stock level are polynomial kernel support vector regression and Gaussian radial basis function kernel support vector regression respectively

- Machine learning model applicable to building stock energy prediction and retrofit energy saving potential evaluation

Abstract

The building sector accounts for 30% of final energy consumption and 28% of global energy-related carbon dioxide emissions, with space heating and cooling consuming a large share of total buildings' energy consumption. Building stock modelling for space heating and cooling energy prediction provides critical insights on the stock energy consumption and aid the building retrofit policy-making process with the evaluation of the energy-saving potential. By combining the physical modelling approach and data-driven approach, a hybrid approach is applicable for modelling the heating and cooling energy consumption of the building stock, including both residential buildings and non-residential buildings. Within this framework, the Urban Modelling Interface (UMI) tool has been used for physical modelling to generate heating and cooling energy use intensity. Then, ten different machine learning models, including Gaussian radial basis function kernel support vector regression, linear kernel support vector regression, polynomial kernel support vector regression, random forests, extreme gradient boosting, ordinary least-squares linear regression, ridge regression, least absolute shrinkage, and selection operator, elastic net and artificial neural network, have been applied to predict heating and cooling energy use intensity (EUI). The approach has been demonstrated using a case study in Chongqing, China. The results show that machine learning models can achieve accurate building heating and cooling EUI prediction, with the polynomial kernel support vector regression showing the best accuracy at the level of a single building, and the Gaussian radial basis function kernel support vector regression performing the best at the stock level. Machine learning models generated by proposed hybrid approach not only provide quickly prediction of building space heating and cooling energy consumption at the stock level, but also support building retrofit decision makings by evaluate energy saving potential of various retrofit options.

Keywords: Building energy consumption; Heating and cooling; Building Stock modelling; Hybrid approach; Machine learning

1. Introduction

Buildings are responsible for 30% of final energy consumption and 28% of global energy-related carbon dioxide emissions in 2018 according to the International Energy Agency *[1]*. Building energy conservation and carbon emission reduction are actively promoted by governmental authorities by leveraging on legislation and policies, such as the Energy Performance of Buildings Directive and the Energy Efficiency Directive in the EU *[2]* and the 13th Five Year Plan in China *[3]*.

Space heating and cooling through mechanical systems are the primary active methods to adjust the building indoor thermal conditions but at the expense of a significant amount of energy. As examples, in residential buildings the space heating and cooling account for 58% and 41% of urban and rural household energy consumption in China

[4], 48% of home energy consumption in the United States *[5]*, 70% of domestic energy consumption in the United Kingdom *[6]* and 65% of the household energy consumption in the European Union *[7]*. In non- residential buildings, the space heating and cooling account for 34% of commercial building energy consumption in the United States *[8]*, 50%-60% of public building energy consumption in China *[9]*, and 45% of non-domestic premises energy consumption across England and Wales *[10]*. The high energy demand for space heating and cooling thus entails massive building energy conservation and carbon emissions reduction potential if tailored building retrofit measures are undertaken.

To understand the building stock energy consumption and study various building retrofit measures, building stock energy modelling - a successor of building energy modelling – is utilized to expand the study area to a larger scale and offers architects, urban planners, and policymakers a valid decision support tool *[11]*. Modelling the space heating and cooling energy consumption boosts policy-making process by providing critical insights on the building stock built environment control-related energy consumption; further, it proves particularly useful to areas in which building energy consumption statistics is lacking, or detailed building end-use split for space heating and cooling is not available. Moreover, the space heating and cooling energy consumption model is also capable of evaluating the energy conservation potential of various building retrofit measures at the stock level and help with the selection of the best performing measures.

This study deployed a hybrid approach to generate data-driven energy prediction model for large-scale building stock covering both residential building and non-residential building without existing building energy consumption data. The structure of the paper is as following: Section 2 includes the related literatures as well as the aims and objectives of this study. Section 3 presents the methodology applied in this study, which use hybrid approach to predict building space heating and cooling energy consumption. Follows by Section 4 demonstrates the proposed hybrid approach using a case study in Chongqing, China. The discussions and conclusions of the study are covered in Section 5 and Section 6 respectively.

2. Literature review

2.1 Data-driven building energy consumption prediction

The data-driven building energy consumption prediction has been gaining raising research interest in recent years *[12]*: it has been widely used to predict building energy consumption of buildings with different functions, such as residential *[13-22]*, office *[23-29]*, institutional *[30, 31]*, educational *[32, 33]* and commercial *[34]*. However, the application of the data-driven approach in large scale building stock energy consumption prediction is rather limited *[34-36]*, this might because the majority of existing research about data-driven building energy consumption prediction is focused on residential or non-residential buildings only *[12]*, although building stock usually consists of a mix of both types of building. Build up a data-driven energy consumption prediction framework able to handle buildings of different functions is essential for extending the application of data-driven approach in large scale building stock.

To the best of our knowledge, there are only a few data-driven building energy consumption prediction studies considering both residential and non-residential buildings, such as that of *Georgescu, et al. [37]* who studied offices, laboratories, gymnasiums, dormitories, and restaurants. Instead of creating one model able to predict both the residential and non-residential building's energy demand, they generated an individual support vector machine model for building energy consumption data from every building utility meters. *Kontokosta and Tull [38]* applied linear regression, random forest, and support vector regression algorithms to predict the energy use of 1.1 million buildings in New York City of various functions, the building energy usage data used to train the model came from Local Law 84 energy disclosure data. *Hawkins, et al. [39]* used the artificial neural network to estimate the energy use in UK university campus buildings, such as dormitories, laboratories, and offices, by using Display Energy Certificate (DEC) to develop artificial neural network energy prediction model. *Robinson, et al. [40]* developed 11 different machine learning models using the Commercial Buildings Energy Consumption Survey (CBECS) data to estimate commercial building energy consumption. The commercial buildings have been studied including both commercial buildings for a residential purpose like lodging building and commercial buildings for non-residential purpose like the office building. Similarly, *Cheng [41]* also based on the CBECS data to build 10 machine learning models for commercial building energy prediction, benchmarking data of New York City and Chicago has been used for model validation. *Abbasabadi, et al. [42]* demonstrated an integrated data-driven framework for urban energy use modelling taking Chicago as a case study. They tested multiple linear regression, nonlinear regression, classification and regression trees, random decision forest, k-nearest neighbours and artificial neural intelligence for operational energy use prediction considering both residential and non-residential buildings. The building energy data used is obtained by merging the Chicago energy benchmark and Chicago energy usage datasets. *Pan and Zhang [43]* employed categorical boosting model, random forest and gradient boosting decision tree in estimate energy consumption of non-residential and multifamily building, Seattle's building energy performance data collected by Seattle's Energy Benchmarking Program is used as main dataset. However, the rich building energy consumption datasets, like Local Law 84 energy disclosure data, DEC data, CBECS data, Chicago energy benchmark dataset and Seattle's building energy performance data, are currently available only for a limited number of cities and countries. The lack of building energy consumption datasets *[44]*, needed as a training set, impede the use of a data-driven approach in the large scale building stock *[45]*.

2.2 Hybrid approach in building stock energy modelling

Top-down and *bottom-up* methods are generally used to develop building stock models *[46-48]*. Top-down methods have embedded the main limitation of lack of technical detail specifications and are unable to determine the energy consumption of each enduses *[46-48]*, while bottom-up methods overcome this shortcoming and are used to investigate the building energy consumption for heating and cooling in this study. Two main approaches for bottom-up building stock energy modelling are typically employed *[46, 47, 49]*: the physical modelling and the data-driven approach. Physical modelling relies on thermodynamic laws for detailed energy modelling, it large data and computational demands stopped it to apply precisely in every building at the stock level *[40]*. The data-driven approach “learns” from historical or available datasets for prediction *[12]*, a large amount of data is essential for model development *[50]*.

The hybrid approach combines physical modelling and data-driven approaches by using the output of physical modelling as an input to generate data-driven models *[40, 50]*. It has the potential to provide a solution for building energy consumption datasets lacking by using physical modelling to generate datasets. Therefore, a hybrid approach has been identified as a more promising method for urban energy modelling *[42]*. *Valovcin, et al. [51]* built multiple linear regressions to adjust energy simulation results to match the measured energy data in U.S. homes as a part of statistical post-processing techniques. Similarly, *Brøgger, et al. [52]*, *[53]* adopted a hybrid approach by using multiple linear regression to calibrate a physical model of the Danish residential building stock. *Li and Yao [54]* compared the performance of linear regression, artificial neural network and support vector regression in predicting the residential annual space heating and cooling loads. The annual residential heating and cooling load intensity database utilized in machine learning models’ training and validation process is generated by EnergyPlus simulation of a typical residential household archetype. *Ciulla and D'Amico [55]* undertook a parametric simulation of a detailed TRNSYS model and generated a building energy database representative of non-residential Italian building stocks. Based on the database, multiple linear regression models are develop to predict building heating, cooling and comprehensive energy demand. *Luo, et al. [56]* proposes a multi-objective prediction framework for building heating, cooling, lighting loads and BIPV electrical power production. By using building operating and energy data generated by TRNSYS simulation of a general office building, artificial neural network, support vector regression and long-short-term-memory neural network based predictive models are trained and tested. Although adapted a hybrid approach, the aforementioned five studies focus on the residential building or non-residential building only. *Goel, et al. [57]* build random forest regression models based on building stock simulations for buildings energy efficiency prediction in developing the Asset Score Preview tool, a rating system tool. In their research, 22 building types embedding both commercial buildings and mid- to high-rise residential buildings were studied with one regression model generated per every building type. There is a lack of study using hybrid approach for energy modelling of both residential building and non-residential building to enable large-scale building stock energy prediction.

2.3 Aims and objectives

To extend the application of data-driven model to large-scale building stock and to alleviate the challenges of commonly unavailable building energy consumption data to support model generation, a hybrid approach has been employed to develop a data-driven energy prediction model covering both residential and non-residential buildings. A case study in Chongqing city (China) is used to demonstrate the hybrid energy prediction approach, the prediction accuracy of ten different machine leaning models is also compared based on the case study.

3. Methodology

The proposal of a new hybrid approach for building energy stock modelling consists of 5 steps, including the heating and cooling energy consumption estimation, machine learning models, model generation process, model performance evaluation as well as the application of selected machine learning model (see Figure 1).

Step 1: Based on building information collected through a field survey and related building characteristics settings, Urban Modeling Interface (UMI) was used to simulate the space heating and cooling energy consumption of all single-use buildings within the study stock.

Step 2: Suitable machine learning models for predicting building space heating and cooling energy use intensity (EUI) at the individual building level have been investigated.

Step 3: Generation of the machine learning models through pre-process of the raw dataset; train with the training and validation set, and test models by apply them to predict the EUIs of the testing set buildings.

Step 4: Evaluate the prediction accuracy of the machine learning models at both individual building and stock levels to compare the machine models' performance when considering both residential and non-residential buildings.

Step 5: Based on the further analysis scope, prioritize building level accuracy or stock level accuracy to select the best performed model. The selected machine learning model can be applied to building space heating and cooling energy consumption prediction, as well as building retrofit space heating and cooling energy saving potential evaluation.

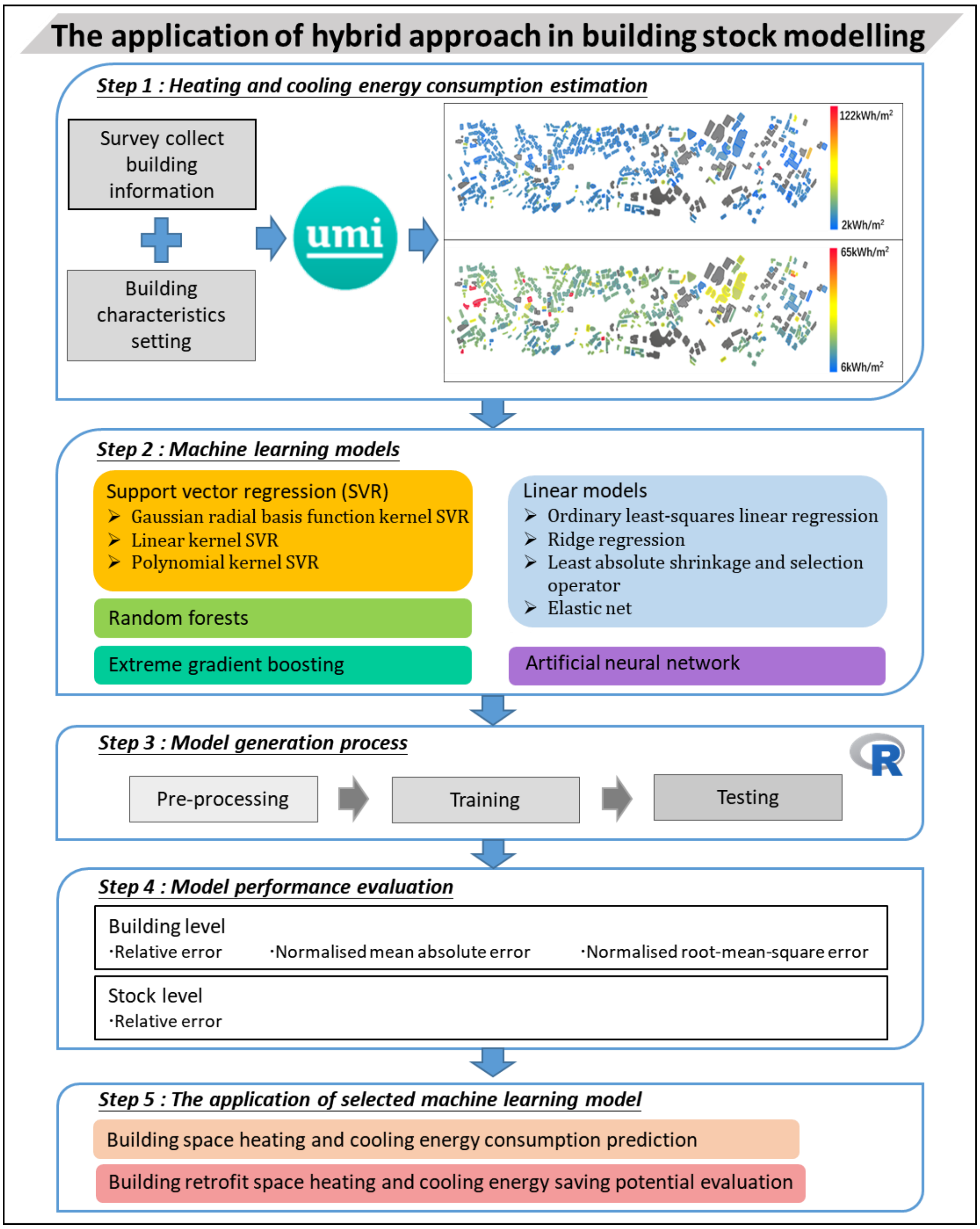

Figure 1: Framework of the research

The detail implication of those five steps is described in the following sections 3.1 to 3.5.

3.1. Heating and cooling energy consumption estimation

As stated above, the rich building energy consumption datasets are not commonly available, so the building energy consumption information needed for data-driven model development is estimated by using physical models. In this study, the energy consumption of every studied building is simulated individually by using Urban Modeling Interface (UMI) *[58]*, a modelling software package that utilizes EnergyPlus *[59]* as the simulation core engine. UMI can simulate space heating and cooling energy use intensity (EUI) for individual buildings at the urban scale in a fast but accurate manner by using a 'shoeboxer' algorithm *[60]*, which makes it a handy physical modelling tool to handle a relatively small scale buildings stock. UMI needs 3D building model of the stock, together with all detailed building characteristics required by EnergyPlus, such as the building envelope thermal physical characteristics and HVAC system, at individual building level to simulate building heating and cooling energy consumption. As detailed building characteristics are essential for UMI simulation, the UMI simulation setting and running process are both labour intensive and time-consuming *[61]*, which does limit its applicability to the large scale building stock.

The heating and cooling energy consumption results from UMI simulation is combined with the building detailed characteristics to create the machine learning database. The database is divided into two subsets and utilised in two ways: 1) as training and validation set to train machine learning models; 2) as testing set to test the performance of machine learning models and compare their accuracy with UMI simulation.

3.2. Machine learning models

Five classes of machine learning technique are investigated in this study to predicting space heating and cooling energy consumption, including support vector regression, random forest, extreme gradient boosting, linear model and artificial neural network. Ten different machine learning models are built based on the machine learning database generated in the previous step.

3.2.1. Support vector machine

Commonly recognized as the best supervised learning algorithms in solving regression, problems *[62]*, SVMs are increasingly used in building energy analysis *[63]*. Introduced by *Cortes and Vapnik [64]* in 1995, the support vector machine (SVM) was initially developed in the context of classification. Based on structural risk minimization inductive principle, SVM aims at minimizing the generalization error through reducing a summation of empirical risk and a Vapnik Chervonenkis (VC) dimension term, which generally leads to higher generalization performance in solving nonlinear problems *[62]*. Support vector regression (SVR), as an extension of the support vector classification (SVC), provides a quantitative response to the input predictor variables *[65]*. It seeks coefficients to minimise the effect of outliers on the regression equations; however, only residuals larger in absolute value than some positive constant(ε) are considered in the loss function *[65, 66]*. ε-insensitive loss functions (equation 1) were used to construct the SVR model and ensure robust and sparse estimation. Only when the discrepancy between the SVR model predicted building EUI and simulated building EUI is higher than ε, the absolute difference will contribute to the loss.

$$L(\mathrm{y}-\mathrm{f}(\mathrm{x}))=\begin{cases}0, if\ |\mathrm{y}-\mathrm{f}(\mathrm{x})|\leq\varepsilon;\\ |\mathrm{y}-\mathrm{f}(\mathrm{x})|-\varepsilon, otherwise.\end{cases} \quad (1)$$

In the case of linear functions $\mathrm{f}(x)=\langle w,x\rangle+b$ with $w\in\mathrm{X}, \mathrm{b}\in P(\langle\cdot,\cdot\rangle$denotes the dot product in X), given training data $\{(x_1,y_1),\ldots,(x_n,y_n)\}\subset X\times P$. The goal of SVR is to find a function $\mathrm{f}(x)$ that has at most ε deviation from the obtained targets for all the training data, and at the same time is as flat as possible. Slack variables ξ_i and ξ_i^* are introduced to guard against outliers and to adopt the soft-margin approach, in case the convex optimization problem is not always feasible. The optimization problem is presented in equation 2 *[67]*.

$$\text{minimize } \frac{1}{2}||w||^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \quad (2)$$

$$\text{subject to } \begin{cases}y_i-\langle w,x\rangle-b\leq\varepsilon+\xi_i\\ \langle w,x\rangle+b-y_i\leq\varepsilon+\xi_i^*\\ \xi_i,\xi_i^*\geq 0\end{cases}$$

C is a positive constant that measures the trade-off between the flatness of function $\mathrm{f}(x)$ and the amount up to which deviations larger than ε are tolerated.

The abovementioned optimization problem can be solved by constructing a Lagrange function, the function $\mathrm{f}(x)$ can be derived as equation 3 *[67]*,

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\langle x_i,x\rangle+b \quad (3)$$

Where, $\alpha,\alpha^*$are Lagrange multipliers of non-negative real numbers.

In the case of nonlinear functions, as the relationship between the building heating/cooling EUI and the selected predictor variables, the predictor variables need to be pre-processed and map from input space into feature space. The function $\mathrm{f}(x)$ is written as equation 4 *[67]*:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\, k(x_i, x) + b \quad (4)$$

Three different kernel functions $k\langle x_i, x\rangle$ is used to generate three different SVR models, including Linear kernel(equation 5) for Linear kernel SVR, Polynomial kernel (equation 6) for polynomial kernel SVR and Gaussian radial basis function kernel (equation 7) for Gaussian radial basis function kernel SVR *[68]*.

$$k(x_i, x) = x_i \cdot x \quad (5)$$

$$k(x_i, x) = (scale \cdot x_i \cdot x + offset)^{degree} \quad (6)$$

$$k(x_i, x) = \exp(-\sigma ||x_i - x||^2) \quad (7)$$

3.2.2. Random forests

Random forests is an ensemble learning approach to supervised learning *[69]*, it can be used for both classification and regression. Thanks to the advantage of fast training speed *[70]*, random forests becomes one of the most widely used machine learning techniques *[71]*. The random forest for regression is formed by growing trees depending on a random vector such that the tree predictor takes on numerical values by average the prediction of every tree *[72]*. The algorithm for random forest regression is as following *[73]*,

1. For b=1 to B:
   (a) Draw a bootstrap sample **Z*** of size N from the training data.
   (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size S_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variables/split-point among the m variables.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees$\{T_b\}_1^B$.

To make a prediction at a new point x:

$$\hat{f}_{rf}^{B}(x) = \frac{\sum_{b=1}^{B} T_b(x)}{B}$$

Where B is the number of trees.

3.2.3. Extreme gradient boosting

Extreme gradient boosting, commonly referred to as XGBoost, is a scalable machine learning system for tree boosting *[74]*. As one of the boosting models, extreme gradient boosting grow trees sequentially. Starting from building the first tree based on the training data, then a second tree is created to correct the errors from the first tree. More trees are added until the model can predict the training set perfectly or the number of trees reaches the upper limit. Extreme gradient boosting is '*an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable*' *[75]*, and can be used to handle regression, classification, and ranking problems *[76]*. Extreme gradient boosting achieved state-of-the-art results in machine learning competitions *[77]*, and was proved to outperform other ten machine learning models at commercial building energy consumption prediction *[40]*.

Based on data set with n examples and m features $D = \{(X_i, y_i)\}$ (|D|=n, $X_i \in P^m$, $y_i \in P$), extreme gradient boosting predicts output by using K additive functions, as shown in equation 8 *[74]*.

$$\hat{y}_i = \emptyset(X_i) = \sum_{k=1}^{K} f_k(X_i)\,, f_k \in \Phi, \ (8)$$

Each f_k corresponds to an independent tree structure, Φ is the space of regression trees.

The regularized objective function presented in equation 9 is optimized in extreme gradient boosting to learn the set of functions *[74]*,

$$\Lambda(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega\,(f_k) \ (9)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$

l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, while Ω is model complexity penalization term. T is the number of leaves in the tree, ω is the leaf weights.

The more detailed mathematical implication of extreme gradient boosting can be found in *Chen and Guestrin [74]* and *Chen and He [78]*.

3.2.4. Linear models

For linear models, the relationship between the predicted variable and predictors can directly or indirectly be written according to the following equation 10 *[66]*. They are selected for their simplicity, intuitive and ability to provide a baseline performance measure *[55, 79]*.

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_j x_{ij} + e_i \quad (10)$$

where y_i is the numeric response for the ith sample; b_0 is the estimated intercept; b_j is the estimated coefficient for the jth predictor variable; x_{ij} is the value of the jth predictor variable for the ith sample; and e_i is the random error of the linear regression model.

For ordinary least-squares linear regression, the aim is to minimise the sum-of-squared errors (SSE_{ols}, shown in equation 11) between the observed value and model-predicted value *[66]*.

$$\mathrm{SSE}_{ols} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (11)$$

The y_i and $\widehat{y_i}$ are the observed value and model-predicted value of the ith sample.

In ridge regression, to pursue smaller mean squared error, a biased model is generated by adding a penalty to the SSE_{rr} *[80]* as shown in equation 12:

$$\mathrm{SSE}_{rr} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda \sum_{i=1}^{n} b_j^2 \quad (12)$$

For the least absolute shrinkage and selection operator model *[81]*, as the SSE_{lasso}(shown in equation 13) is penalized by the absolute values, the penalty value λ can reach 0, so the lasso model also conducts feature selection.

$$\mathrm{SSE}_{lasso} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda \sum_{i=1}^{n}\left|b_j\right| \quad (13)$$

The elastic net model combined two types of penalties to enable effective regularization via the ridge-type penalty with the feature selection quality of the lasso penalty*[66]*.

The SSE_{en} is presented in the following equation 14 *[82]*:

$$SSE_{en} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{i=1}^{n} b_j^2 + \lambda_2 \sum_{i=1}^{n} \left|b_j\right| \quad (14)$$

3.2.5. Artificial neural network

With the benefits of high speed, high accuracy, and capability of handling nonlinear relationships between variables *[83]*, artificial neural network is the most widely applied artificial intelligence models in the building energy prediction *[63]*. It mimics how the brain responds to stimuli from sensory inputs to interpret the relationship between input and output signals *[84]*. The neuron is the information-processing unit of the neural network, the mathematical description of a neuron is shown in equation 15 *[85]*:

$$y_k = \varphi\left(\sum_{j=1}^{m} w_{kj} x_j + b_k\right) \quad (15)$$

where, $x_1, x_2, \ldots x_m$ are the input signals; $w_{k1}, w_{k2}, \ldots w_{km}$ are the synaptic weights of neuron k; b_k is the bias; $\varphi(\cdot)$ is the activation function; and y_k is the output signal of the neuron.

3.3. Model generation process

Machine learning models are generated via the process presented in Figure 2. All predictor variables are centred and scaled as pre-process before model training to avoid domination from attributes in higher numeric range and improve numerical stability *[24, 66]*. After the pre-processing, all the available data are randomly divided into two parts, with 25% as the testing set and 75% as the training and validation set (the residential building and non-residential building ratio remain equal in both datasets), as the 25/75 split is commonly used in machine learning related studies *[54, 86-88]*. Then, all data in the training and validation set is further partitioned into ten equally sized subsets and undergo the 10-fold cross-validation process. By repeating the process of using nine subsets as a training set and one subset as the validation set for 10 times, the tuning parameter(s) of the machine learning models are determined as the one(s) with the best average performance for the 10 different validation sets. Then, the final model is generated using all data from the training and validation set and the untouched testing set is used to evaluate the prediction accuracy of the models.

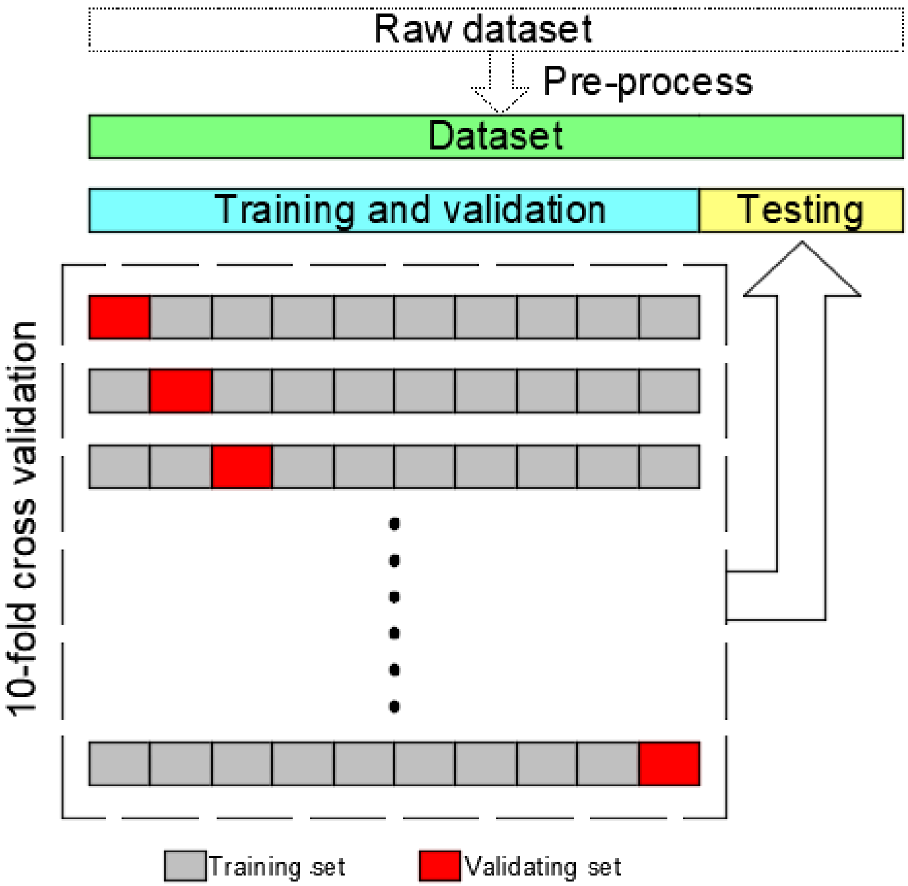

Figure 2: Machine learning model generation process

3.4. *Model performance evaluation*

All buildings in the testing set are used to evaluate the performance of the machine learning model in predicting EUI as an unseen dataset. The accuracy of the machine learning-based model on individual building heating and cooling EUI prediction is investigated using relative error as per equation 16:

$$\delta_k = \frac{\widehat{y_k} - y_k}{y_k} \times 100\% \quad (16)$$

Here δ_k is the relative error of using ‘machine learning’-based model to predict heating/cooling EUI of building k against UMI simulations;

y_k is the building heating/cooling EUI for building k from the UMI simulation generated database;

$\widehat{y_k}$ is the predicted building heating/cooling EUI for building k from the machine learning model;

The average prediction performance of different machine learning models at the individual building level is indicted by normalised mean absolute error (NMAE) and normalised root-mean-square error (NRMSE) for heating and cooling EUI. Their calculation formulas are presented in equations 17-18.

$$\mathrm{NMAE} = \frac{\frac{\sum_{k=1}^{n} |y_k - \widehat{y_k}|}{n}}{\frac{\sum_{k=1}^{n} y_k}{n}} \quad (17)$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{k=1}^{n} (y_k - \widehat{y_k})^2}{n}}}{\frac{\sum_{k=1}^{n} y_k}{n}} \quad (18)$$

Where n is the total number of buildings in the testing set.

To evaluate the accuracy of machine learning models on whole stock, residential stock and non-residential stock level energy prediction, the relative error of gross heating and cooling energy consumption of all buildings in the testing set, all residential buildings in the testing set and all non-residential buildings in the testing set are estimated using equation 19 respectively:

$$\delta_{Stock} = \frac{\sum_{k=1}^{m}(\widehat{y_k} \times F_k) - \sum_{k=1}^{m}(y_k \times F_k)}{\sum_{k=1}^{m}(y_k \times F_k)} \quad (19)$$

Where, δ_{Stock} is the relative error of using machine learning based models to predict the gross heating/cooling energy consumption of specific building stock in the testing set; m is the total number of buildings in the testing set belongs to the specific building stock; F_k is the total floor area of the building k.

Apart from the prediction accuracy indexes described above, the running time to predict the heating and cooling EUI of all buildings in the testing set is also tracked and analysed.

3.5. The application of selected machine learning model

By comparing the prediction accuracy indexes of all ten machine learning models, the best performed model can be selected based on the further analysis scope. If predicting the space heating and cooling energy consumption precisely in the building level is more important, then the building level accuracy indexes should be prioritize. Otherwise, the best performed model should be select based on the stock level accuracy indexes. The selected machine leaning model is applicable to predict building space heating and cooling energy consumption, evaluate energy saving potential for retrofit measures as a substitute of building physical simulation.

4. Case study

The case study area is located in Yuzhong District of Chongqing city (China), covering an area of about 3.4 km^2 (see Figure 3). From July 2015 to September 2015, a field survey was carried out to collect detailed building information for every building within the study area; collected information included buildings' geographic location (longitude and latitude), function, construction age, number of floors, window-to-wall ratio. For construction age, instead of specific construction completed year, age band was collected. Including three age bands for residential buildings (pre-2001, 2001-2010 and post-2010) and four age bands for non-residential buildings (Pre-1990, 1990-2005, 2005-2015 and Post-2015). The construction age are collected for the building construction information plaque and by asking the owners.

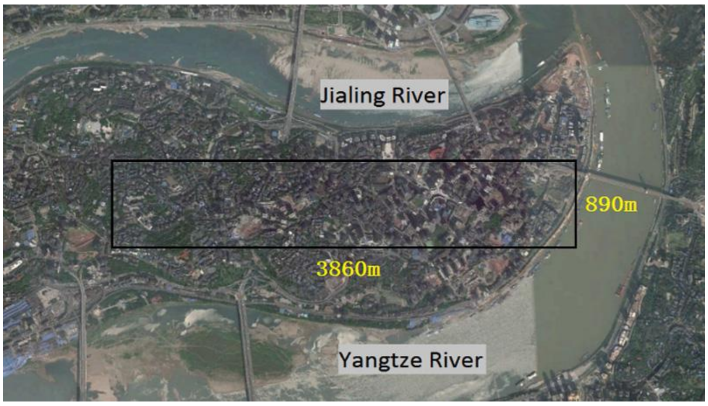

Figure 3: The case study area (highlighted by a black box) within the Yuzhong district

The geographic location is used for locating buildings on online maps, then a building stock 3D model is generated by extrude the footprints by its height. The height of every building is calculated using the following equation 20, while the window-to-wall ratio is set according to the filed survey.

$$D=N \times d \quad (20)$$

Where D is the building height; N is the number of floors the building have; d is the average floor height, according to standards, it was set as three meters for residential buildings *[89]*, four meters for offices *[90]*, educational buildings *[91, 92]*, hospitals *[93]* and hotels *[94]*, five meters for malls *[95]*.

4.1. Characteristics of the buildings in the study area

In total, there are 573 buildings located within the case study area. One hundred thirty-one of which are mixed-use multifunction building, while the rest of them are hosting a single function (including educational buildings, hospital, hotel, mall, office, and residential buildings). The specific location of each building in the study area is shown in Figure 4.

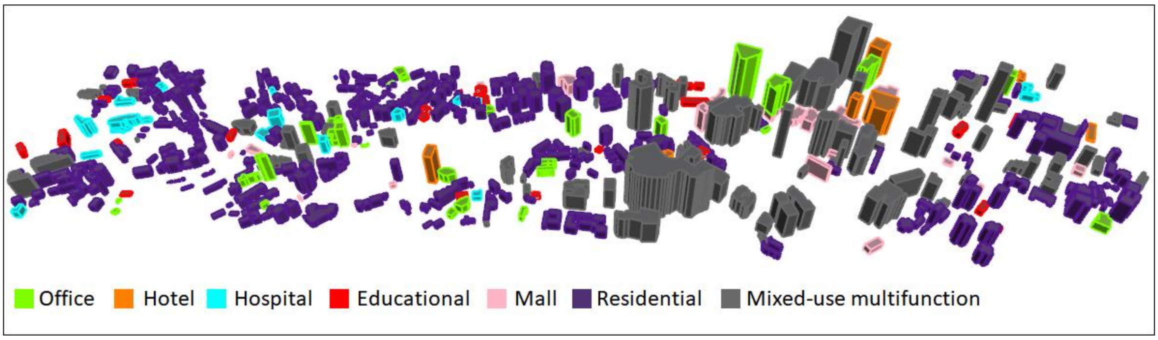

Figure 4: Building location and function

The total number of single functions buildings is presented in Figure 5, including 334 residential buildings and 108 non-residential buildings. The residential building is dominating the case study area as it accounted for more than three-quarters of all single-function buildings. The construction age distribution of residential and non-residential buildings is presented in Figure 6, majority of residential buildings are constructed before 2001, while more than half of non-residential buildings are constructed during 1990 to 2005.

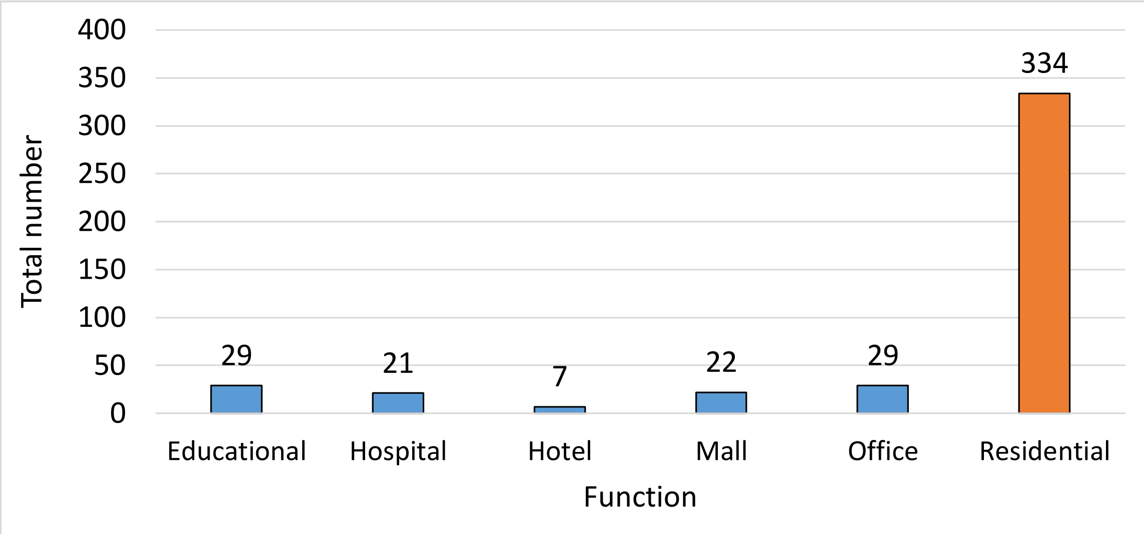

Figure 5: The total number of buildings with different functions

Figure 6: The construction age distribution of residential and non-residential buildings

In this study, only the 442 single function buildings are studied, due to the difficulty in getting the real floor area function within mixed-use buildings.

The building's characteristics, including thermo-physical characteristics of the building envelope, HVAC systems, and internal loads, are set according to the Chinese national and industrial design standards based on the construction age of the buildings. JGJ 134-2001 *[96]*, and JGJ 134-2010 *[97]* Standards are utilized to describe the building characteristics of the residential building of different construction age. GBJ 19-1987 *[98]*, GB 50189-2005 *[99]* and GB 50189-2015 *[100]* Standards are used to describe the characteristics of non-residential buildings. The detailed building characteristics setting for the residential and non-residential building is set according to *Costanzo, et al. [101]*, and are shown in Table 1.

Table 1: Detailed building characteristics of non-residential and residential building *[101]*

Building function		Construction age	Building envelope thermal-physical characteristics					HVAC system			Internal gains		
			U-values (W/m²K)				Infiltrations (ACH)	Fresh air supply (m³/s· p)	Heating/Cooling setpoint (°C)	Heating efficiency/Cooling EER (-)	Occupants density (p/m²)	Equipment density (W/m²)	Lighting density (W/m²)
			Walls	Roof	Slab	Windows (U value/SHGC)							
Non-residential building	Office	Pre-1990	1.95	1.44	3.79	5.74/0.85	0.25	0.005	20/26	0.55/3.8	0.25	20	11
		1990-2005	1.44	0.97	1.88	5.74/0.85	0.25	0.005	20/26	0.55/3.8	0.25	20	11
		2005-2015	0.95	0.78	0.97	2.67/0.43	0.15	0.008	20/26	0.89/4.1	0.25	20	11
		Post-2015	0.5	0.69	0.7	2.50/0.34	0.15	0.008	20/26	0.9/4.8	0.1	15	9
	Hotel	Pre-1990	1.95	1.44	3.79	5.74/0.85	0.25	0.008	20/26	0.55/3.8	0.067	20	11
		1990-2005	1.44	0.97	1.88	5.74/0.85	0.25	0.008	20/26	0.55/3.8	0.067	20	11
		2005-2015	0.95	0.78	0.97	2.67/0.43	0.15	0.008	20/26	0.89/4.1	0.067	20	11
		Post-2015	0.5	0.69	0.7	2.50/0.34	0.15	0.008	20/26	0.9/4.8	0.04	15	7
	Mall	Pre-1990	1.95	1.44	3.79	5.74/0.85	0.25	0.002	20/26	0.55/3.8	0.33	13	12
		1990-2005	1.44	0.97	1.88	5.74/0.85	0.25	0.008	20/26	0.55/3.8	0.33	13	12
		2005-2015	0.95	0.78	0.97	2.67/0.43	0.15	0.005	20/26	0.89/4.1	0.33	13	12
		Post-2015	0.5	0.69	0.7	2.50/0.34	0.15	0.008	20/26	0.9/4.8	0.125	13	10
	Hospital	Pre-1990	1.95	1.44	3.79	5.74/0.85	0.25	0.004	20/26	0.55/3.8	0.125	20	15
		1990-2005	1.44	0.97	1.88	5.74/0.85	0.25	0.004	20/26	0.55/3.8	0.125	20	15
		2005-2015	0.95	0.78	0.97	2.67/0.43	0.15	0.008	20/26	0.89/4.1	0.125	15	12
		Post-2015	0.5	0.69	0.7	2.50/0.34	0.15	0.008	20/26	0.9/4.8	0.125	15	8

	Educational	Pre-1990	1.95	1.44	3.79	5.74/0.85	0.25	0.005	20/26	0.55/3.8	0.25	20	11
		1990-2005	1.44	0.97	1.88	5.74/0.85	0.25	0.005	20/26	0.55/3.8	0.25	20	11
		2005-2015	0.95	0.78	0.97	2.67/0.43	0.15	0.008	20/26	0.89/4.1	0.25	20	11
		Post-2015	0.5	0.69	0.7	2.50/0.34	0.15	0.008	20/26	0.9/4.8	0.17	5	9
Residential building		Pre-2001	1.97	1.62	3.74	5.74/0.85	2	0	18/26	1/2.2	0.03	4.3	6
		2001-2010	1.03	1	1.5	2.80/0.48	1	0	18/26	1.9/2.3	0.03	4.3	6
		Post-2010	0.83	0.8	1.31	2.67/0.34	1	0	18/26	1.9/2.3	0.03	4.3	6

473

For non-residential buildings, the HVAC system is supposed to be in use for the whole year, from 7 AM to 7 PM (12h) every weekday for office and educational buildings; 24h every day for hotel and hospital building; 8 AM-10 PM (14h) every day for the mall. The HVAC system is available for the heating period (from December 1st to February 28th) and cooling period (from June 15th to August 31st) only for residential buildings. The daily residential HVAC usage is assumed based on the study of *Hu, et al. [102]*, as an hour in the morning (from 7 AM-8 AM) and five hours when returning home from work (from 6 PM-11 PM) for heating, as well as 6 PM-8 AM (14 hours) and 1 PM-2 PM (1 hour) for cooling.

4.2. Buildings' energy consumption

The results of the UMI simulations are presented in Figure 7, heating and cooling EUIs are available at the individual building level. As shown in Figure 8, heating EUI varies from 2 kWh/m^2 to 122 kWh/m^2, while the cooling EUI varies from 6 kWh/m^2 to 65 kWh/m^2 for all 442 single function buildings studied. The building energy consumption data is combined with building detailed characteristics to create the database used to develop machine learning models.

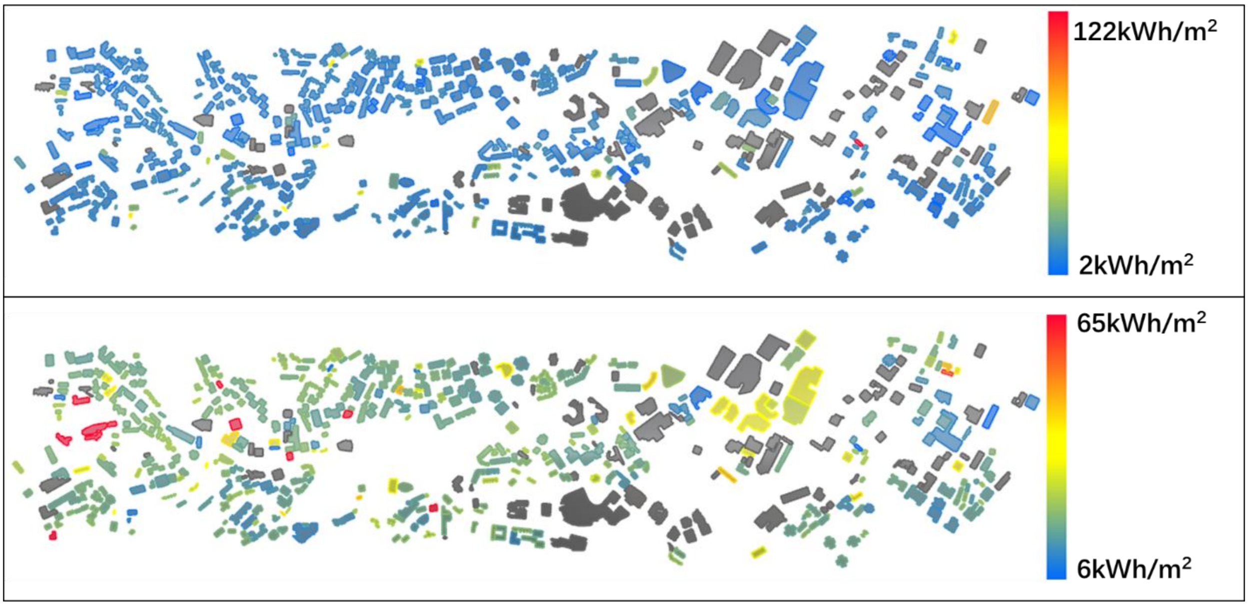

Figure 7: The heating (top) and cooling (bottom) EUI of buildings in the study area (the buildings fill in grey are mixed-use buildings which are not simulated)

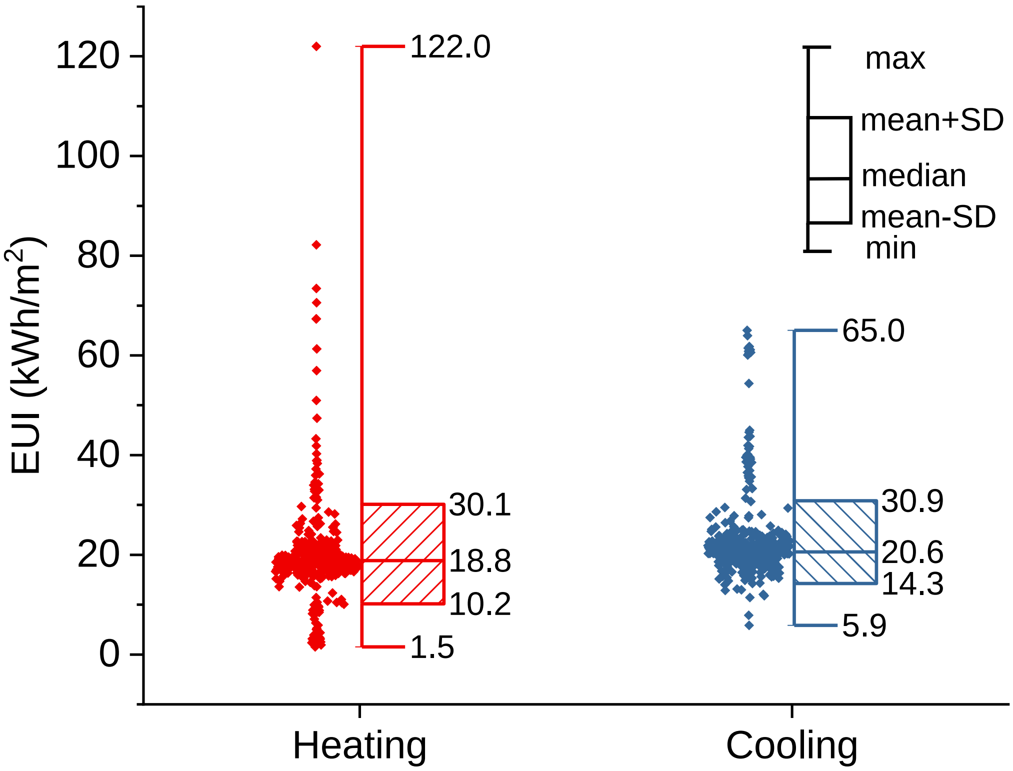

Figure 8: Boxplots of heating and cooling EUIs

4.3. Predictor variables selection

The building characteristics (listed in Table 2), including building geometry, building envelope thermal-physical characteristics, building HVAC system and building internal gains, are considered as main predictor variables as they are the main determinants for building space heating and cooling energy consumption *[103]*. Predictor variables of building geometry, building envelope thermo-physical characteristics, and building internal gains are considered for both heating and cooling EUI prediction, while the selection of predictor variables for building HVAC system is different. For heating EUI prediction, only the fresh air supply, heating temperature setpoint, the heating efficiency, and heating available proportion are considered, likewise, for cooling EUI correlation analysis, only the fresh air supply, the cooling COP and cooling available proportion are considered. The cooling setpoint is excluded from being a predictor variable because of its constant value of 26 °C for all buildings.

Table 2: Predictor variables for heating and cooling EUI prediction [orange shading marks those used for heating EUI prediction only; blue shading marks those used for cooling EUI prediction only; unshaded ones are used for both heating and cooling EUI prediction]

Building characteristics	Predictor variables
Building geometry	Building height [m]
	Compactness ratio [/]
	Window to wall ratio [/]
Building envelope thermal-physical characteristics	Walls U-value [W/m^2K]
	Roof U-value [W/m^2K]
	Slab U-value [W/m^2K]
	Windows U-value [W/m^2K]
	Windows solar heat gain coefficient (SHGC) [/]
	Air infiltrations [ach]
Building HVAC system	Fresh air supply [$m^3/s{\cdot}p$]
	Heating setpoint [°C]
	Heating efficiency [/]
	HVAC available proportion for heating [/]
	Cooling EER [/]
	HVAC available proportion for cooling [/]
Building internal gains	Occupants density [p/m^2]
	equipment density [W/m^2]
	Lighting density [W/m^2]

The compactness ratio (*CR*) is an index of building shape, and is calculated as per following Equation 21 *[61]*:

$$CR=S/V \quad (21)$$

Where S is the surface area of the building;

V is the enclosed volume of the building.

The HVAC available proportion (AP) for heating and cooling indicated the annual portion of time when the HVAC system is available for heating and cooling respectively; they are calculated using Equation 22:

$$AP=H/8760 \quad (22)$$

Where H is the total number of hours per annual when heating (or cooling) is available from the HVAC system.

4.4. Prediction accuracy analysis

The caret package*[104]* developed by Max Kuhn for predictive model generating has been used to perform all the machine learning models under R programming language. Caret was set to automatically generate 5 values for each tuning parameter, the tuning parameters combination with the best accuracy in the training and validation set is used in the final model for prediction accuracy analysis. As the 110 buildings in the testing set are not used for training of the machine learning models, the prediction accuracy in the testing set can reasonably represent the prediction accuracy of applying those machine learning models to other single-function buildings in Chongqing.

The relative error distribution of applying machine-learning models in heating and cooling EUI for all buildings in the testing set is shown in Figure 9. The machine learning models give an accurate prediction about building heating and cooling EUI. The percentage of building within the ±10% relative error varies between 61.8% (ordinary least-squares linear regression and least absolute shrinkage and selection operator) to 85.5% (polynomial kernel support vector regression), and from 81.8% (linear kernel support vector regression) to 91.8% (Gaussian radial basis function kernel support vector regression) for the heating and cooling cases, respectively. The percentage of building within the ±20% relative error varies between 80.0% (ridge regression and elastic net) to 90.9% (polynomial kernel support vector regression) and

94.5% (linear kernel support vector regression and ordinary least-squares linear regression) to 98.2% (artificial neural network) for heating and cooling.

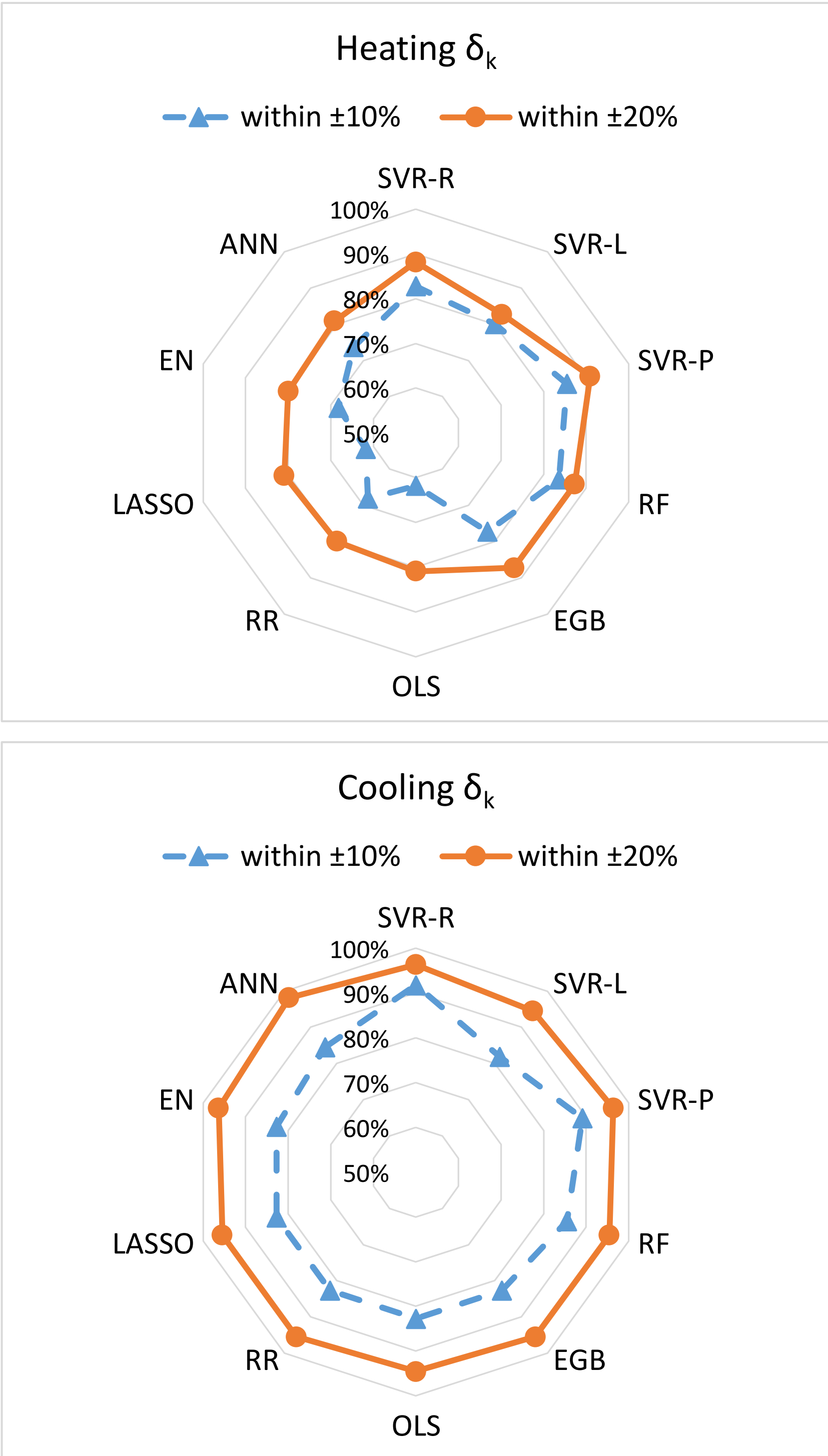

Figure 9: The relative error of the machine learning models in building heating (top) and cooling (bottom) EUI prediction (SVR-R: Gaussian radial basis function kernel support vector regression; SVR-L: linear kernel support vector regression; SVR-P:

polynomial kernel support vector regression; RF: random forests; EGB: extreme gradient boosting; OLS: ordinary least-squares linear regression; RR: ridge regression; LASSO: least absolute shrinkage and selection operator; EN: elastic net; ANN: artificial neural network)

The NMAE and NRMSE of applying different machine learning models in the testing set are presented in Table 3. In general, the prediction accuracy for cooling EUI is better than heating EUI with smaller NMAE and NRMSE. For heating EUI prediction, the NMAE varies from 7.3% (polynomial kernel SVR) to 17.2% (both ordinary least-squares linear regression and least absolute shrinkage and selection operator model), the NRMSE varies from 17.3% (polynomial kernel SVR) to 46.2% (extreme gradient boosting). For cooling EUI prediction, the NMAE varies from 4.3% (polynomial kernel SVR) to 6.4% (both linear kernel SVR and elastic net), the NRMSE varies from 6.2% (polynomial kernel SVR) to 13.4% (extreme gradient boosting). The polynomial kernel SVR has the best accuracy in the individual building level, followed by Gaussian radial basis function kernel SVR.

Table 3: NMAE and NRMSE results of different machine learning models

Machine learning models	**Heating EUI**		**Cooling EUI**	
	NMAE	**NRMSE**	**NMAE**	**NRMSE**
Gaussian radial basis function kernel SVR	8.4%	19.3%	4.9%	8.8%
Linear kernel SVR	11.2%	23.9%	6.4%	10.9%
Polynomial kernel SVR	7.3%	17.3%	4.3%	6.2%
Random forests	12.0%	40.1%	5.2%	8.9%
Extreme gradient boosting	13.3%	46.2%	6.0%	13.4%
Ordinary least-squares linear regression	17.2%	35.7%	6.2%	10.3%
Ridge regression	15.8%	32.1%	6.3%	9.7%
Least absolute shrinkage and selection operator	17.2%	35.7%	5.9%	9.3%
Elastic net	15.8%	32.1%	6.4%	10.1%
Artificial neural network	13.8%	29.8%	6.1%	8.9%

The performance of machine learning models in stock level heating and cooling energy consumption prediction is presented in Table 4. For the whole stock including both residential and non-residential building, the relative error for heating and cooling at the whole stock level are within ±4%, except for heating prediction of artificial neural network which has a relative error of -9.7%. Heating energy consumption is more likely to be underestimated, with cooling energy consumption are more likely to be overestimated. The Gaussian radial basis function kernel SVR performed the best with a whole stock level relative error of -0.2% and -0.3% respectively for heating and cooling prediction. Followed by polynomial kernel SVR, with a whole stock level relative error of 0.3% and 0.5% respectively for heating and cooling prediction. It is interesting to note that although the artificial neural network has a high relative error for heating prediction, it performs very well in cooling prediction with a relative error of only 0.2%. For the residential stock, random forests and extreme gradient boosting performed the best in heating and cooling prediction respectively, with relative error of 0.6% and 0.1%. For the non-residential stock, linear kernel SVR and polynomial kernel SVR performed the best in heating and cooling prediction respectively, with relative error of 2.0% and -1.0%. Meanwhile, all machine learning models studied overestimate space cooling energy consumption for residential stock while underestimate space cooling energy consumption for non-residential stock

Table 4: The relative error δ_{Stock} of different machine learning models at the stock level

Machine learning models	**Whole stock**		**Residential stock**		**Non-residential stock**	
	Heating	**Cooling**	**Heating**	**Cooling**	**Heating**	**Cooling**
Gaussian radial basis function kernel SVR	-0.2%	-0.3%	1.7%	0.6%	-6.2%	-2.4%
Linear kernel SVR	1.0%	1.1%	0.7%	2.2%	2.0%	-1.7%
Polynomial kernel SVR	0.3%	0.5%	2.9%	1.1%	-7.8%	-1.0%
Random forests	-0.6%	-0.8%	0.6%	1.0%	-4.4%	-5.0%
Extreme gradient boosting	2.4%	-1.0%	5.5%	0.1%	-7.7%	-3.8%
Ordinary least-squares linear regression	-3.8%	1.7%	-7.6%	3.2%	8.2%	-2.1%
Ridge regression	-2.3%	0.5%	-5.7%	1.4%	8.8%	-1.6%
Least absolute shrinkage and selection operator	-3.8%	0.8%	-7.6%	2.0%	8.2%	-2.3%
Elastic net	-2.3%	0.6%	-5.7%	1.8%	8.8%	-2.6%
Artificial neural network	-9.7%	0.2%	-9.0%	0.7%	-11.7%	-1.2%

The running time of applying machine learning models in building heating and cooling EUI prediction is shows in Figure 10, varies from 0.032 seconds for elastic net to 0.769 seconds for extreme gradient boosting. All ten machine learning models studied are able to predict the heating and cooling EUI of 110 buildings within 1 second, while using UMI to simulation heating and cooling EUI of one building takes at least 10 seconds. The machine learning models can speed up the building heating and cooling EUI prediction for more than 1000 times, the swift speed benefits the large scale building stock energy prediction by greatly reduce the prediction time it takes. The machine learning models' running time and UMI simulation time presented above are based on a ThinkPad personal computer with Intel Core i7-6500U Processor, 8 GB

RAM, and Windows 10 64-bit operating system. The times may vary when using a different computer.

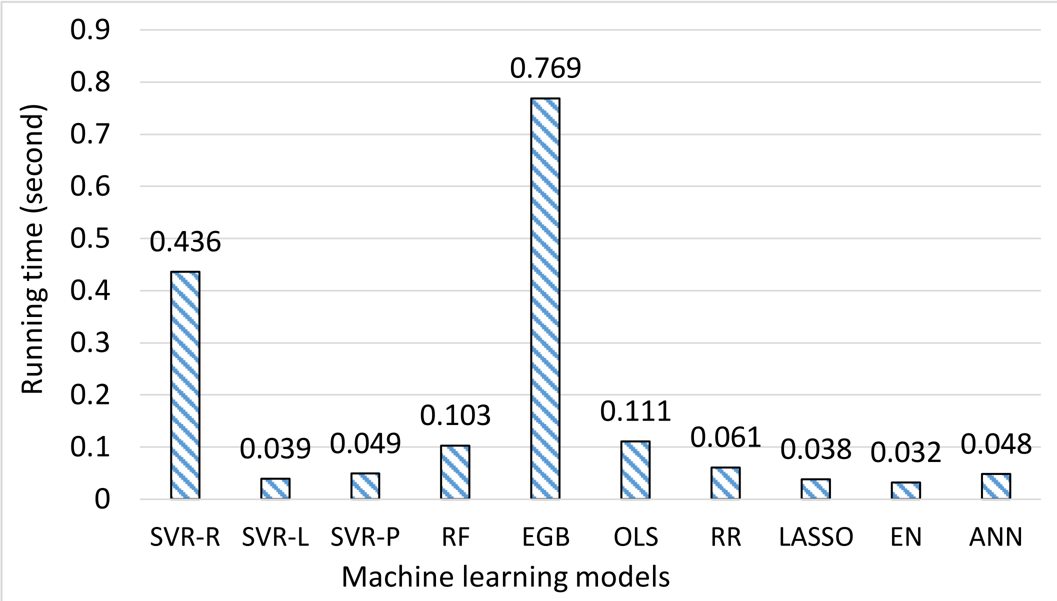

Figure 10: The running time of applying machine learning models in building heating and cooling EUI prediction of 110 buildings in testing set(SVR-R: Gaussian radial basis function kernel support vector regression; SVR-L: linear kernel support vector regression; SVR-P: polynomial kernel support vector regression; RF: random forests; EGB: extreme gradient boosting; OLS: ordinary least-squares linear regression; RR: ridge regression; LASSO: least absolute shrinkage and selection operator; EN: elastic net; ANN: artificial neural network)

4.5. Evaluation of building stock retrofit energy saving potential

This section demonstrates the application of machine learning model in building stock retrofit energy saving potential evaluation. As Gaussian radial basis function kernel SVR performed the best at the whole stock level, it is utilized to show the energy saving potential of upgrading building envelopes for entire stock. Assuming to improve the building thermo-physical performance by ensure all buildings' envelope meet the latest standard. The building envelope thermo-physical characteristics for older buildings, including pre-2015 non-residential buildings and pre-2010 residential buildings, after retrofit are shown in Table 5.

Table 5: Assumed building envelope thermal-physical characteristics after retrofit

Building function	**Construction age**	**Building envelope thermal-physical characteristics**				
		U-values (W/m^2K)				**Infiltrations (ACH)**
		Walls	Roof	Slab	Windows (U value/SHGC)	
Non-residential building	Pre-1990	0.5	0.69	0.7	2.50/0.34	0.15
	1990-2005	0.5	0.69	0.7	2.50/0.34	0.15
	2005-2015	0.5	0.69	0.7	2.50/0.34	0.15
Residential building	Pre-2001	0.83	0.8	1.31	2.67/0.34	1
	2001-2010	0.83	0.8	1.31	2.67/0.34	1

The gross space heating and cooling energy consumption figures for all the buildings in the testing set before and after retrofit are shown in Figure 11. By improving the building envelope, energy consumption reduction is achieved in both space cooling and space heating, with the latter showing a more substantially decrease. The building retrofit performance evaluation using Gaussian radial basis function kernel SVR is straightforward, use the updated U-values and infiltration rates together with other predictor variables which stay unchanged, a swift estimation of the building space heating and cooling demand after retrofit can be achieved. Compared to re-run UMI simulation with updated building envelope thermal-physical characteristics, the machine learning model is faster and less computation intensive.

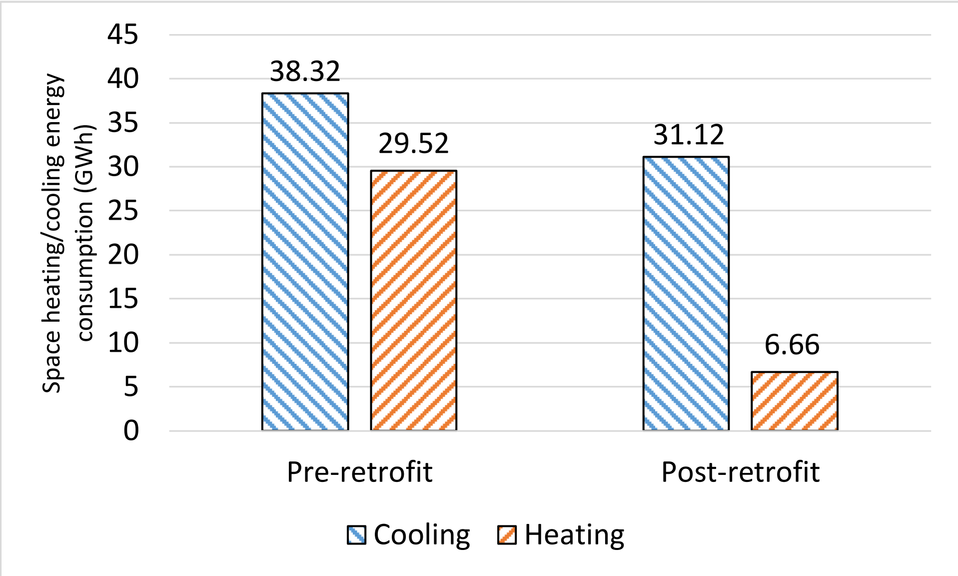

Figure 11: The gross space heating and cooling energy consumption before and after retrofit

5. Discussions and limitations

Starting from training and validation set generated by UMI small scale building stock dynamic simulation, machine learning models are developed via pre-processing and training. The performances of the machine learning models are tested using the UMI generated test set, the comparison shows that machine learning models can replace UMI to predict the heating and cooling energy consumption of single-function buildings in Chongqing with accuracy. Moreover, their swift running time enables potential large-scale building stock energy consumption prediction. The hybrid approach proposed in this study provide a way to give an insight view of the space heating and cooling energy consumption of residential and non-residential building at a large scale building stock. Which helps the understanding of the current energy used in adjust the building indoor thermal conditions. This provides a solid start point for energy conservation related policy making when the real space heating and cooling energy consumption data is not available due to reasons like lack of monitoring. As detailed building characteristics are used as the predictor variables of the machine learning model, the energy-saving potential of various building retrofit options can be evaluated by the machine learning model. The identification of the best performed retrofit option can support policy making about large scale building stock energy conservation. Moreover, machine learning modelling is easy to use even for people without great knowledge about building thermal physics, so will also be a handy tool for the general public to evaluate the retrofit energy-saving potential of various retrofit options.

Although the hybrid approach proposed in this study can predict the building space heating and cooling energy consumption, the lack of public available building energy consumption datasets in Chongqing hinders the validation and calibration of the model to real building energy consumption. The collection of real building energy consumption data remains as a very important task to understand and bridge the performance gap between predicted energy use and actual energy use *[105]*. Moreover, collecting other building characteristics information, including construction type, construction material, HVAC system, retrofit history record, etc., is also very important in give a true building profile and support building energy consumption calibration.

This study also bears the limitation of considering only the single-function buildings, future works should be carried on to collect detail floor area function information and develop data-driven building energy consumption approach for mixed-use buildings.

6. Conclusions

This study investigated the process of utilizing a hybrid approach to predict building space heating and cooling energy consumption for both residential and non-residential buildings to support large scale building stock energy modelling. Considering the commonly building energy data lacking, the hybrid approach has been used to combine the advantages of both physical modelling and data-driven approaches.

Based on the building energy consumption data generated by UMI physical modelling, ten different data-driven machine learning models, including Gaussian radial basis function kernel SVR, linear kernel SVR, polynomial kernel SVR, random forests, extreme gradient boosting, ordinary least-squares linear regression, ridge regression, least absolute shrinkage, and selection operator, elastic net and artificial neural network, have been trained to predict heating and cooling energy use intensity for both residential buildings and non-residential buildings (containing educational buildings, hospitals, hotels, malls, and offices). Building characteristics are utilized as predictor variables of those machine learning models, including geometry characteristics, envelope thermal-physical characteristics, HVAC system characteristics and internal gains characteristics. With known predictor variables, the machine leaning models are able to predict building heating and cooling energy use intensity at individual building level. A case study in Chongqing city (China) has been used to demonstrate the proposed process and test the prediction accuracy of machine learning models. The main findings are summarized as follows:

- Machine learning models can handle both residential and non-residential building energy consumption prediction using a single model, so there is no need to generate multiple models according to different building functions.
- Machine learning models can accurately predict building heating and cooling EUI, with polynomial kernel support vector regression, predicted 85.5% of building heating EUI within ±10% of relative error and Gaussian radial basis function kernel support vector regression predicted 91.8% of building cooling EUI within ±10% of relative error.
- The polynomial kernel SVR has the best accuracy in the individual building level, with NMAE and NRMSE for heating EUI as 7.3% and 17.3% respectively; NMAE and NRMSE for cooling EUI as 4.3% and 6.2% respectively.
- The Gaussian radial basis function kernel SVR performed the best in the whole stock level, with a relative error of only -0.2% and -0.3% respectively for heating and cooling prediction.
- Use machine learning models for building heating and cooling energy consumption prediction is more than 1000 times faster than UMI physical modelling, their swift speed proved their potential in large-scale building stock

energy modelling.

By integrating physical modelling with data-driven machine learning techniques, the hybrid approach for modelling heating and cooling energy consumption of building stock is no longer rely on the availability of building energy consumption data. Moreover, it can speed up the process of building stock modelling by decrease the number of buildings to be physically simulated and dramatically cutting down the processing time. The generated machine learning model can be applied to quickly predict building space heating and cooling energy consumption at the stock level, as well as evaluate energy saving potential of different building stock retrofit options. This is of great help for building energy conservation related decision makings, it not only provide an insight view of the current space heating and cooling energy consumption when the monitored data is not available, but also able to compare various retrofit measures and select the best one to be implicated in the whole stock. Although the hybrid approach is only demonstrated in Chongqing in this paper, it can be easily replicated in other cities and countries.

Acknowledgements

This research is financially supported by the Natural Science Foundation of China (grant numbers NSFC 51561135002) and the UK Engineering and Physical Sciences Research Council (grant number EPSRC EP/N009797/1) for the collaborative China-UK project LoHCool Low carbon climate-responsive Heating and Cooling of cities.

The authors would like to thank all who participated in the field survey and the comments from Yuxin Wu (Chongqing University) and Vincenzo Costanzo (University of Catania).

References

[1] IEA. Buildings-Tracking Clean Energy Progress. 2019; https://www.iea.org/tcep/buildings/ (Accessed 2019.03.25)

[2] European Commission. Buildings. 2019; https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings (Accessed 2019.03.25)

[3] MOHURD. Building Energy Conservation and Green Building Development 13th Five Year Plan. 2017; http://www.mohurd.gov.cn/wjfb/201703/t20170314_230978.html (Accessed 2017.3.23)

[4] X. Zheng, C. Wei, P. Qin, J. Guo, Y. Yu, F. Song, Z. Chen, Characteristics of residential energy consumption in China: Findings from a household survey, Energy Policy, 75 (2014) 126-135.

[5] EIA. Heating and cooling no longer majority of U.S. home energy use. 2013; https://www.eia.gov/todayinenergy/detail.php?id=10271&src=%E2%80%B9%20Consumption%20%20%20%20%20%20Residential%20Energy%20Consumption%20Survey%20(RECS)-f4 (Accessed 2018.2.18)

[6] Department for Business Energy & Industrial Strategy, Energy consumption in the UK 2017, in, 2017.

[7] Eurostat. Energy consumption in households. 2018; http://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_consumption_in_households#cite_note-1 (Accessed 2018.3.2)

[8] EIA. Energy Use in Commercial Buildings. 2018; https://www.eia.gov/energyexplained/index.php?page=us_energy_commercial#tab1 (Accessed 2019.03.22)

[9] M. LI, Influence of Indoor Air Computation Parameter of Civil Building to Heating and Air-Conditioning Energy Consumption, Master, Tianjin University, 2010.

[10] BEIS. Building Energy Efficiency Survey (BEES). 2013;

https://www.gov.uk/government/publications/building-energy-efficiency-survey-bees (Accessed 2019.03.24)

[11] C.F. Reinhart, C.C. Davila, Urban building energy modeling - A review of a nascent field, Building and Environment, 97 (2016) 196-202.

[12] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews, 81 (2018) 1192-1205.

[13] J.-S. Chou, D.-K. Bui, Modeling heating and cooling loads by artificial intelligence for energy-efficient building design, Energy and Buildings, 82 (2014) 437-446.

[14] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Applied Energy, 123 (2014) 168-178.

[15] F. Lai, F. Magoulès, F. Lherminier, Vapnik's learning theory applied to energy consumption forecasts in residential buildings, International Journal of Computer Mathematics, 85 (10) (2008) 1563-1588.

[16] J. Ma, J.C.P. Cheng, Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests, Applied Energy, 183 (2016) 193-201.

[17] J. Ma, J.C.P. Cheng, Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology, Applied Energy, 183 (2016) 182-192.

[18] E. Mocanu, P.H. Nguyen, M. Gibescu, W.L. Kling, Deep learning for estimating building energy consumption, Sustainable Energy, Grids and Networks, 6 (2016) 91-99.

[19] S. Naji, A. Keivani, S. Shamshirband, U.J. Alengaram, M.Z. Jumaat, Z. Mansor, M. Lee, Estimating building energy consumption using extreme learning machine method, Energy, 97 (2016) 506-516.

[20] S. Paudel, M. Elmitri, S. Couturier, P.H. Nguyen, R. Kamphuis, B. Lacarrière, O. Le Corre, A relevant data selection method for energy consumption prediction of low energy building based on support vector machine, Energy and Buildings, 138 (2017) 240-256.

[21] L. Wei, W. Tian, E.A. Silva, R. Choudhary, Q. Meng, S. Yang, Comparative Study on Machine Learning for Urban Building Energy Analysis, Procedia Engineering, 121 (2015) 285-292.

[22] U. Ali, M.H. Shamsi, M. Bohacek, K. Purcell, C. Hoare, E. Mangina, J. O'Donnell,

A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making, Applied Energy, 279 (2020) 115834.

[23] H. Deng, D. Fannon, M.J. Eckelman, Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata, Energy and Buildings, 163 (2018) 34-43.

[24] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings, 37 (5) (2005) 545-553.

[25] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Applied Energy, 86 (10) (2009) 2249-2256.

[26] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks, Energy Conversion and Management, 50 (1) (2009) 90-96.

[27] S.L. Wong, K.K.W. Wan, T.N.T. Lam, Artificial neural networks for energy analysis of office buildings with daylighting, Applied Energy, 87 (2) (2010) 551-557.

[28] H.X. Zhao, F. Magoulès, Parallel support vector machines applied to the prediction of multiple buildings energy consumption, Journal of Algorithms & Computational Technology, 4 (2) (2010) 231-249.

[29] T. Liu, Z. Tan, C. Xu, H. Chen, Z. Li, Study on deep reinforcement learning techniques for building energy consumption forecasting, Energy and Buildings, 208 (2020) 109675.

[30] Z. Wang, Y. Wang, R.S. Srinivasan, A novel ensemble learning approach to support building energy use prediction, Energy and Buildings, 159 (2018) 109-122.

[31] F. Zhang, C. Deb, S.E. Lee, J. Yang, K.W. Shah, Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique, Energy and Buildings, 126 (2016) 94-103.

[32] C. Fan, F. Xiao, Y. Zhao, A short-term building cooling load prediction method using deep learning algorithms, Applied Energy, 195 (2017) 222-233.

[33] R. Wang, S. Lu, W. Feng, A novel improved model for building energy consumption prediction based on model integration, Applied Energy, 262 (2020).

[34] S. Walker, W. Khan, K. Katic, W. Maassen, W. Zeiler, Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings, Energy and Buildings, 209 (2020) 109705.

[35] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy

consumption, Renewable and Sustainable Energy Reviews, 82 (2018) 1027-1047.

[36] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review, Energy and Buildings, 165 (2018) 301-320.

[37] M. Georgescu, E. Eccles, V. Manjunath, E. Swindle, I. Mezi, Machine learning methods for site-level building energy forecasting and data rectification, in: Building Simulation and Optimization–The Second IBPSA–England Conference, 2014.

[38] C.E. Kontokosta, C. Tull, A data-driven predictive model of city-scale energy use in buildings, Applied Energy, 197 (2017) 303-317.

[39] D. Hawkins, S.M. Hong, R. Raslan, D. Mumovic, S. Hanna, Determinants of energy use in UK higher education buildings using statistical and artificial neural network methods, International Journal of Sustainable Built Environment, 1 (1) (2012) 50-63.

[40] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M.A. Brown, R.M. Pendyala, Machine learning approaches for estimating commercial building energy consumption, Applied Energy, 208 (2017) 889-904.

[41] X. Cheng, Applying Machine Learning Based Data-Driven Approach in Commercial Building Energy Prediction, ASHRAE Transactions, 126 (2020) 403-411.

[42] N. Abbasabadi, M. Ashayeri, R. Azari, B. Stephens, M. Heidarinejad, An integrated data-driven framework for urban energy use modeling (UEUM), Applied Energy, 253 (2019) 113550.

[43] Y. Pan, L. Zhang, Data-driven estimation of building energy consumption with multi-source heterogeneous data, Applied Energy, 268 (2020).

[44] H. Rashid, P. Singh, A. Singh, I-BLEND, a campus-scale commercial and residential buildings electrical energy dataset, Sci Data, 6 (2019) 190015.

[45] S. Fathi, R. Srinivasan, A. Fenner, S. Fathi, Machine learning applications in urban building energy performance forecasting: A systematic review, Renewable and Sustainable Energy Reviews, 133 (2020).

[46] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Building and Environment, 45 (7) (2010) 1683-1697.

[47] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, Renewable and Sustainable Energy Reviews, 13 (8) (2009) 1819-1835.

[48] M. Brøgger, K.B. Wittchen, Estimating the energy-saving potential in national building stocks – A methodology review, Renewable and Sustainable Energy Reviews, 82 (2018) 1489-1496.

[49] H. Lim, Z.J. Zhai, Review on stochastic modeling methods for building stock energy prediction, Building Simulation, 10 (5) (2017) 607-624.

[50] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renewable and Sustainable Energy Reviews, 23 (2013) 272-288.

[51] S. Valovcin, A.S. Hering, B. Polly, M. Heaney, A statistical approach for post-processing residential building energy simulation output, Energy and Buildings, 85 (2014) 165-179.

[52] M. Brøgger, P. Bacher, K.B. Wittchen, A hybrid modelling method for improving estimates of the average energy-saving potential of a building stock, Energy and Buildings, 199 (2019) 287-296.

[53] M. Brøgger, P. Bacher, H. Madsen, K.B. Wittchen, Estimating the influence of rebound effects on the energy-saving potential in building stocks, Energy and Buildings, 181 (2018) 62-74.

[54] X. Li, R. Yao, A machine-learning-based approach to predict residential annual space heating and cooling loads considering occupant behaviour, Energy, 212 (2020) 118676.

[55] G. Ciulla, A. D'Amico, Building energy performance forecasting: A multiple linear regression approach, Applied Energy, 253 (2019) 113500.

[56] X.J. Luo, L.O. Oyedele, A.O. Ajayi, O.O. Akinade, Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads, Sustainable Cities and Society, 61 (2020) 102283.

[57] S. Goel, H. Horsey, N. Wang, J. Gonzalez, N. Long, K. Fleming, Streamlining building energy efficiency assessment through integration of uncertainty analysis and full scale energy simulations, Energy and Buildings, 176 (2018) 45-57.

[58] Sustainable Design Lab, Umi, in, 2017.

[59] DOE. EnergyPlus Energy Simulation Software. 2017; http://apps1.eere.energy.gov/buildings/energyplus/ (Accessed 2017.12.14)

[60] T. Dogan, C. Reinhart, Shoeboxer: An algorithm for abstracted rapid multi-zone urban building energy model generation and simulation, Energy and Buildings, 140 (2017) 140-153.

[61] X. Li, R. Yao, M. Liu, V. Costanzo, W. Yu, W. Wang, A. Short, B. Li, Developing urban residential reference buildings using clustering analysis of satellite images,

Energy and Buildings, 169 (2018) 417-429.

[62] F. Magoulès, H.-X. Zhao, Data Mining and Machine Learning in Building Energy Analysis: Towards High Performance Computing, John Wiley & Sons, 2016.

[63] H.-x. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews, 16 (6) (2012) 3586-3592.

[64] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning, 20 (3) (1995) 273-297.

[65] R. Tibshirani, G. James, D. Witten, T. Hastie, An introduction to statistical learning-with applications in R, in, New York, NY: Springer, 2013.

[66] M. Kuhn, K. Johnson, Applied predictive modeling, Springer, 2013.

[67] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing, 14 (3) (2004) 199-222.

[68] K. Hornik, D. Meyer, A. Karatzoglou, Support vector machines in R, Journal of statistical software, 15 (9) (2006) 1-28.

[69] R.I. Kabacoff, R in Action, Second Edition ed., manning, New York, 2015.

[70] Y. Zhu, W. Xu, G. Luo, H. Wang, J. Yang, W. Lu, Random Forest enhancement using improved Artificial Fish Swarm for the medial knee contact force prediction, Artificial Intelligence in Medicine, 103 (2020) 101811.

[71] A.A. Ahmed Gassar, G.Y. Yun, S. Kim, Data-driven approach to prediction of residential energy consumption at urban scales in London, Energy, 187 (2019).

[72] L. Breiman, Random Forests, Machine Learning, 45 (1) (2001) 5-32.

[73] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, The Mathematical Intelligencer, 27 (2) (2005) 83-85.

[74] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.

[75] XGBoost Documentation. 2020; https://xgboost.readthedocs.io/en/latest/ (Accessed 2020.06.29)

[76] T. Chen, T. He, M. Benesty. Xgboost presentation. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboostPresentation.html (Accessed 2020.06.29)

[77] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling

the energy consumption of commercial buildings, Energy and Buildings, 158 (2018) 1533-1543.

[78] T. Chen, T. He, Higgs boson discovery with boosted trees, in: NIPS 2014 workshop on high-energy physics and machine learning, 2015, pp. 69-80.

[79] R.E. Edwards, J. New, L.E. Parker, Predicting future hourly residential electrical consumption: A machine learning case study, Energy and Buildings, 49 (2012) 591-603.

[80] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12 (1) (1970) 55-67.

[81] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1) (1996) 267-288.

[82] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2) (2005) 301-320.

[83] Z. Li, J. Dai, H. Chen, B. Lin, An ANN-based fast building energy consumption prediction method for complex architectural form at the early design stage, Building Simulation, 12 (4) (2019) 665-681.

[84] B. Lantz, Machine learning with R, Packt Publishing Ltd, 2015.

[85] S. Haykin, Neural Networks: A Comprehensive Foundation, Second Edition, Pearson Education, India, 1999.

[86] T. Han, A. Siddique, K. Khayat, J. Huang, A. Kumar, An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete, Construction and Building Materials, 244 (2020).

[87] D. Geysen, O. De Somer, C. Johansson, J. Brage, D. Vanhoudt, Operational thermal load forecasting in district heating networks using machine learning and expert advice, Energy and Buildings, 162 (2018) 144-153.

[88] D. Assouline, N. Mohajeri, J.-L. Scartezzini, Quantifying rooftop photovoltaic solar energy potential: A machine learning approach, Solar Energy, 141 (2017) 278-296.

[89] MOHURD, Code for design of residential buildings GB 50096-2011, in, Ministry of Housing and Urban-Rural Development, People's Republic of China., 2011.

[90] MOHURD, Design code for office building JGJ67-2006, in, 2006.

[91] MOHURD, Code for design of school GB50099-2011, in, 2011.

[92] MOHURD, Code for design of nursery and kindergarten buildings JGJ39-2016, in, 2016.

[93] MOHURD, Code for design of general hospital GB51039-2014, in, 2014.

[94] MOHURD, Code for Design of Hotel Building JGJ 62-2014, in, 2014.

[95] MOHURD, Code for design of store buildings JGJ48-2014, in, 2014.

[96] MOHURD, Design standard for energy efficiency of residential buildings in hot summer and cold winter zone JGJ 134-2001 (in Chinese), in, Ministry of Housing and Urban-Rural Development, People's Republic of China., 2001.

[97] MOHURD, Design standard for energy efficiency of residential buildings in hot summer and cold winter zone JGJ 134-2010 (in Chinese), in, Ministry of Housing and Urban-Rural Development, People's Republic of China., 2010.

[98] MOHURD, Design code for heating, ventilation and air conditioning GBJ 19-87, in, 1987.

[99] MOHURD, Design standard for energy efficiency of public buildings GB50189-2005, in, Ministry of Housing and Urban-Rural Development, People's Republic of China., 2005.

[100] MOHURD, Design standard for energy efficiency of public buildings GB50189-2015, in, Ministry of Housing and Urban-Rural Development, People's Republic of China., 2015.

[101] V. Costanzo, R. Yao, X. Li, M. Liu, B. Li, A multi-layer approach for estimating the energy use intensity on an urban scale, Cities, 95 (2019) 102467.

[102] T. Hu, H. Yoshino, Z. Jiang, Analysis on urban residential energy consumption of Hot Summer & Cold Winter Zone in China, Sustainable Cities and Society, 6 (2013) 85-91.

[103] Y. Lu, Practical Heating and Air Conditioning Design Manual (second edition), China Architecture & Building Press, Beijing, 2008.

[104] M. Kuhn. The caret Package. 2017; http://topepo.github.io/caret/index.html (Accessed 2018.4.17)

[105] P. de Wilde, The gap between predicted and measured energy performance of buildings: A framework for investigation, Automation in Construction, 41 (2014) 40-49.