Building Energy Efficiency Estimations with Random Forest for Single and Multi-Zones

[1]Ammar Alammar, [2]Abdulrahman Alymani, [3]Wassim Jabi
[1]King Saud University,[2]Alfaisal University, [3]Cardiff University,
[1]{aammar}@ksu.edu.sa, [2]{abalymani}@alfaisal.edu, [3]{jabiw}@cardiff.ac.uk

Surrogate models (SM) present an opportunity for rapid assessment of a building's performance, surpassing the pace of simulation-based methods. Setting up a simulation for a single concept involves defining numerous parameters, disrupting the architect's creative flow due to extended simulation run times. Therefore, this research explores integrating building energy analysis with advanced machine learning techniques to predict heating and cooling loads (KWh/m2) for single and multi-zones in buildings. To generate the dataset, the study adopts a parametric generative workflow, building upon Chou and Bui's (2014) methodology. This dataset encompasses multiple building forms, each with unique topological connections and attributes, ensuring a thorough analysis across varied building scenarios. These scenarios undergo thermal simulation to generate data for machine learning analysis. The study primarily utilizes Random Forest (RF) as a new technique to estimate the heating and cooling loads in buildings, a critical factor in building energy efficiency. Following that, A random search approach is utilized to optimize the hyperparameters, enhancing the robustness and accuracy of the machine learning models employed later in the research. The RF algorithms demonstrate high performance in predicting heating and cooling loads (KWh/m2), contributing to enhanced building energy efficiency. The study underscores the potential of machine learning in optimizing building designs for energy efficiency.

Keywords: *Heating and Cooling loads, Topology, Machine learning, Random Forest*

INTRODUCTION

Buildings account for roughly 40% of global primary energy consumption, with space cooling and heating alone contributing to half of this energy use. Hence, it is crucial to reduce the energy consumption of buildings and mitigate their impact on the environment. Machine learning has become a promising solution for improving the accuracy of energy consumption prediction utilising models such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Random Forest (RF). SVMs provide a method for minimising risk in a structured way that is suitable for situations with few data samples. On the other hand, ANNs are particularly effective since they could learn on their own and can adapt to nonlinear patterns. Although both techniques show potential, SVMs face issues in terms of scalability, while ANNs face the possibility of data loss and the complexities of optimising network design. Random Forest (RF) functions as an ensemble algorithm. Ensemble algorithms are a type of machine learning approaches that combine various learning courses to get accurate predictions Dietterich (2000). RF algorithm can process both numerical and categorical data as well as its training simplicity and ability to generalize over a huge dataset.

The research methodology consists of four key stages designed to meet the research

objectives: Building Generator, Energy Analysis, Regression Modeling and Training, and Cooling and Heating Predictions. The surrogate model is built using Random Forest algorithms. The creation of this surrogate model is elaborated upon through three primary phases: (1) data pre-processing, (2) model training and hyperparameter optimization, and (3) model validation Westermann and Evins (2019). The primary goal of the research was to develop a robust RF model capable of accurately predicting the heating and cooling loads in buildings. This involved creating a comprehensive dataset through parametric generative workflows, conducting thermal simulations, and performing extensive model training and validation using k-fold cross-validation and hyperparameter tuning.

RELATED WORK

Building energy efficiency has attracted significant attention due to its potential to reduce greenhouse gas emissions. A critical review by Seyedzadeh et al. (2018) underscores the importance of integrating energy efficiency measures at the initial stages of building design and in retrofitting existing structures. The paper emphasizes the promising role of artificial intelligence (AI) and machine learning (ML) techniques, such as artificial neural networks, support vector machines, Gaussian-based regressions, and clustering, in forecasting and enhancing building energy performance. In addition, several researchers have introduced Random Forest (RF) for predicting building energy consumption. RF stands out due to its capability to handle both numerical and categorical data, its straightforward training process, and its capacity to generalize across large datasets (Tso and Yau 2007; Yu et al. 2010; Ahmad et al. 2017; Alammar et al. 2021; Alammar and Jabi 2023)

Wang (2018) contributes to this discourse by examining the efficacy of a RF homogeneous ensemble approach for hourly energy prediction in educational buildings. The study highlights the resilience of RF to variations in the number of variables, showcasing its superiority over regression tree and Support Vector Regression models. It also notes the importance of seasonal operational differences, suggesting that energy prediction models could be refined by incorporating data on energy behavior changes across semesters.

Further Rashidifar (2020) introduces a machine learning framework to assess the impact of eight input variables on the heating and cooling loads of residential buildings. Using statistical tools to identify key input variables and the application of both logistic regression and random forest regression, the study demonstrates machine learning's capability to provide accurate predictions of building parameters.

Liu et al. (2021) focuses on predicting building energy consumption based on the design parameters of the building envelope, such as heat transfer and solar radiation absorption coefficients. By integrating data from Revit and DesignBuilder software and employing a Random Forest model, the research underscores the significant impact of certain design parameters on energy consumption. The RF model's superiority over BP-ANN and SVM in prediction accuracy.

Sarmas et al. (2023)) addresses the critical need for predicting the energy savings of energy efficiency renovation actions to optimize financial resource allocation. The paper introduces a machine-learning framework that employs tree-based algorithms with bagging, boosting, and an additional ensemble level to minimize prediction uncertainty.

Together, these studies underscore the transformative potential of machine learning and artificial intelligence in advancing the field of building energy efficiency through enhancing predictive accuracy and informing optimal design and retrofitting strategies.

METHODOLOGY

This study employs a structured approach, combining building energy analysis with advanced machine learning model to predict cooling and heating demands (KWh/m^2). The methodology comprises four stages employed to achieve the research objectives as follow: Building Generation, Energy Analysis, Regression Modeling and Training, and Cooling and Heating Prediction.

Phase 1: Building Generation: In this phase, 12 types of buildings are created through two primary tasks. The first step involves defining utility functions that specify the important characteristics of each building topology, such as the orientation of its faces (North, South, East, West), in terms of solar exposure and energy efficiency. Additionally, when exporting building models, two versions of each model are created. One version includes the glazing percentage to account for natural light and thermal performance, while the other version excludes the glazing percentage. This allows for a comparative investigation of the impact of glazing on building performance and aesthetics. It is crucial to emphasise that the building generator produces two parallel representations of buildings. A Cell is a model that consists of a single zone, whereas a CellComplex is a cellular structure that encompasses many zones using the TopologicPy 3D modelling tool by .Jabi et al. (2018).

Phase 2: Energy Analysis: In this phase, energy simulations are conducted for different datasets. Cooling simulations are initially performed for the Cell, then followed by heating simulations. Then, the same sequence is applied to the CellComplex (Jabi (2022).

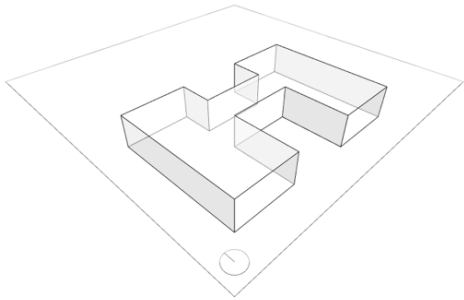
Phase 3: Regression Modeling and Training: Random Forest model was employed to predict the cooling and heating loads (KWh/m^2) of the generated buildings. The model undergoes training to achieve the accurate model. A k-fold cross validation approach was utilized to optimize the hyperparameters, enhancing the robustness and accuracy of the machine learning model. In addition, various metrics are employed to evaluate the model's predictive accuracy and performance comparison.

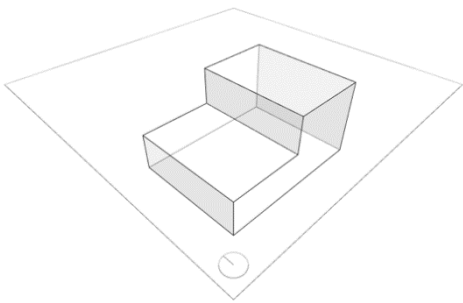
Phase 4: Cooling and Heating Prediction using RF: in this phase the created RF model was tested to predict the cooling and heating loads. (KWh/m^2) of buildings.

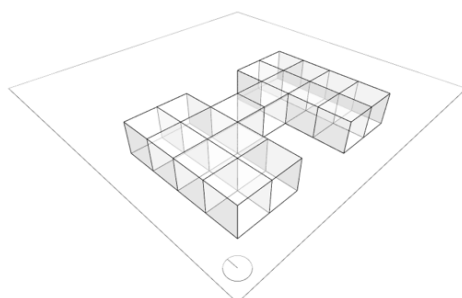
CASE STUDY MODELLING AND SIMULATION

Buildings Modelling and Database Generation

Adhering closely to the configurations specified by Chou and Bui (2014), A 3.5 × 3.5 × 3.5 standard cube was utilized to create 12 distinct architectural designs. Each design consisted of 18 individual components and was generated using the TopologicPy tool within the Jupyter notebook environment. Regardless of differences in surface areas, all constructed buildings maintained a consistent volume of 771.75 m3. Two datasets were generated, each containing 769 structures. The first dataset focuses on a single-zone building, whereas the second dataset includes 18 interconnected cellular spaces referred to as the 'CellComplex', as shown in Figure 2. To create diversity in the level of transparency of the enclosures, the amount and distribution of glazing were altered on different walls and in different orientations. The study analysed three different levels of glazing area: 10%, 25%, and 40% of the total floor area. In addition, the shapes experienced four rotations at 90-degree intervals in the north, east, west, and south directions. A combination of 12 shapes, 3 glazing levels, 5 distribution possibilities, and 4 orientations resulted in a total of 720 simulation variants.







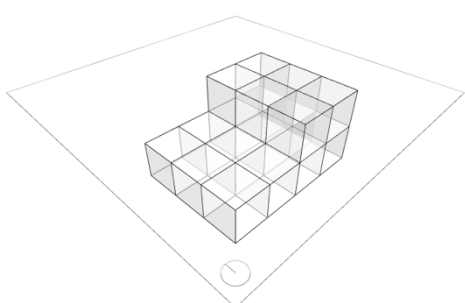


Figure 1 A single space within the building, referred to as the 'cell'.

Figure 2 The architectural configuration comprising 18 spaces within the building, collectively referred to as the 'Cell Complex'.

Thermal Simulation Settings

The materials used for all 18 elements are consistent across different types of buildings. These materials are selected based on their common application in construction and their low U-values. The U-values for different building components are as follows: walls (1.780), floors (0.860), roofs (0.500), and windows (2.260). In line with the prior research conducted by Chou and Bui (2014) the simulation assumes that the buildings are located in Athens, Greece, and employs similar settings for the current study. The architectural configuration comprising 18 spaces within the building, collectively referred to as the 'Cell Complex'.

The simulation began with examining the cooling dataset for the single zone building unit to understand the components that influence its cooling requirements. Afterwards, an examination of the heating dataset was conducted for the same building. This simultaneous technique offered a thorough perspective on the energy demands within a cell. The study further extended its scope to include multi-zoned buildings, which refer to more intricate architectural configurations. The initial simulations focused on analysing cooling data for these buildings, specifically investigating the impact of building interactions on cooling load requirements.

RANDOM FOREST MODEL DEVELOPMENT

As discussed earlier, Random Forest (RF) is an evolution of Decision Trees (DT) which utilized in this experiment for regression learning to model heating and cooling loads (KWh/m^2). The energy data obtained through simulation were utilized to construct the RF model for training and validation. In this study, four experiments were conducted analyzing both the single cell area and the cell complex.

Inputs Features and Data Pre-processing

During testing, the RF model was trained using eight input parameters as variables. The inputs were subjected to pre-processing to improve the accuracy of predictions using machine learning methods. The inputs were classified as either categorical or continuous. Orientation, Glazing Area, and Glazing Distribution were considered as categorical variables and encoded using one-hot encoding. In contrast, the variables Total Height, Relative Compactness, Surface Area, Wall Area, and Roof Area were treated as continuous inputs without any prior data transformation. Each feature showed different ranges that altered the number of building iterations in the models. The model's output includes both cooling and heating loads (KWh/m^2). The network's performance was evaluated using root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R2) score. RMSE is a measure of the average squared difference between the actual and predicted cooling load values. On the other hand, MAE reflects the average absolute difference or residual. Smaller numbers indicate superior model performance. The R^2-value measures the distribution of predicted values around the regression line. It is also known as the coefficient of determination, which indicates the proportion of variance that is accounted for by the model.

K-Fold Cross Validation

In all four experiments, the dataset was divided into training, validation, and testing sets, with 80% allocated to training, 6.67% to validation, and the remaining 13.33% to testing. The training data underwent a 5-fold cross-validation. In k-fold cross-validation, one-fold is designated as the testing set, while the remaining folds are used as the training set. notably, a validation subset of one-third of the testing data was allocated for each validation iteration. RF model accuracy is influenced through various hyperparameters. Therefore, this study considered parameters such as 1) number of trees, (2) bootstrap, and (3) minimal cost-complexity pruning parameter.

EXPERIMENT RESULTS

This section presents the results of the RF model for all experiments after tuning the hyperparameters; **experiment (1):** single cell (cooling loads), **experiment (2):** single cell (heating loads), **experiment (3):** cell complex (cooling loads), **experiment (4):** cell complex (heating loads).

Experiment (1): Cell (cooling loads):

The results from the k-fold cross-validation are depicted in Figure 3. This visualization displays the average RMSE, MAE, and R2 score corresponding to each parameter combination. On the x-axis, different bootstrap settings are presented, with their performance outcomes across varying tree numbers shown as bars. Key findings from the experiments include:

- The difference in performance between enabled and disabled bootstrap options is minor; however, disabling bootstrap leads to decreased performance.
- Performance generally improves with an increased number of trees.
- The optimal model configuration identified from the experiments includes the bootstrap option enabled with a total of 90 trees.

Figure 4, presents the performance metrics of the top-performing models based on the number of trees, all with the bootstrap option enabled. The final test outcomes were: RMSE of 0.050551, MAE of 0.171947, and an R2 score of 0.99177. Figure 5, showcases a visualization of RF predictions for randomly selected energy data. The graph illustrates that, for most data points,

actual and predicted values closely align or are identical.

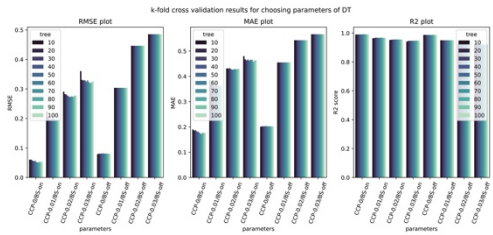


Figure 3 Results of hyperparameter tuning in RF (BS represents the bootstrap option).

	Tree	RMSE	MAE	R^2- score
0	10.0	0.060561	0.190282	0.990126
1	20.0	0.060500	0.185863	0.990150
2	30.0	0.057517	0.187008	0.990655
3	40.0	0.080428	0.201971	0.987150
4	50.0	0.055405	0.181896	0.991005
5	60.0	0.055873	0.181157	0.990864
6	70.0	0.052140	0.176423	0.991508
7	80.0	0.081087	0.202292	0.987054
8	90.0	0.050551	0.171947	0.991776
9	100.0	0.052826	0.176437	0.991366

Table 1 Results of k-fold cross validation for RF.

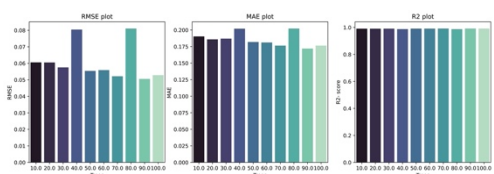


Figure 4 Metrics showcasing the performance of the top-performing models.

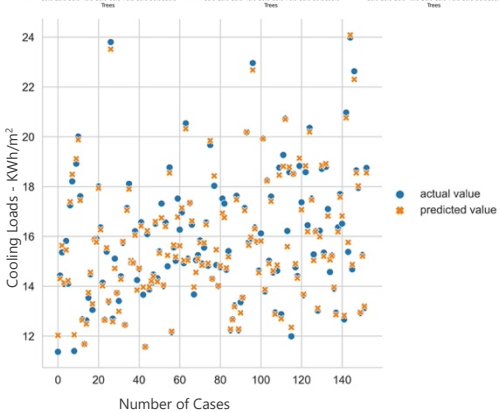


Figure 5 Actual versus predicted values for randomly selected test cases.

Experiment (2): Cell (heating loads):

The k-fold cross-validation outcomes are displayed in Figure 6. From these experiments, several key findings emerged:

- RMSE values exhibited variability with increasing tree numbers, as depicted in.
- The optimal results occurred when the ccp-alpha parameter was set to 0.0, regardless of the number of trees or bootstrap option.
- The performance significantly declined when a non-zero ccp-alpha value was used.
- Models that had the bootstrap option enabled outperformed those that did not.
- The best result was achieved when the number of trees was 100, the ccp-alpha value was 0, and the bootstrap option was enabled Table 2.

The results demonstrate a consistent enhancement in performance as the number of trees increases. Nevertheless, there was a noticeable decrease in the model's performance when 40 and 80 trees were employed. Figure 7, shows the performance metrics of the most successful models, which were determined by the number of trees used and with the bootstrap option enabled. The k-fold cross-validation achieved optimal results while using 100 trees, a ccp-alpha value of 0, and using the bootstrap option. The test results are as follows: the root mean square error (RMSE) is 0.059805, the mean absolute error (MAE) is 0.188470, and the coefficient of determination (R2 score) is 0.990059. The results indicate the model's performance after conducting cross-validation using the entire training dataset and testing it on a 20% sample. Figure 8, displays a graphical representation of the RF predictions for a randomly chosen sample of energy data points. The graph indicates a strong correlation between the actual and predicted values for most data

points, with a high degree of similarity or precision.

Figure 6 Results of hyperparameter tuning in RF.

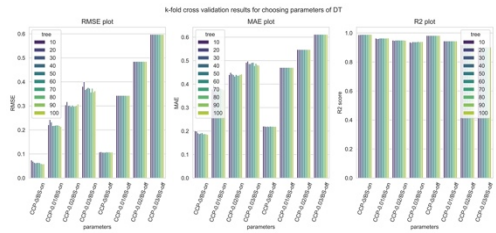


Table 2 The results of k-fold cross validation for RF.

	Tree	RMSE	MAE	R^2- score
0	10.0	0.074548	0.200931	0.987763
1	20.0	0.069566	0.199699	0.988710
2	30.0	0.064396	0.191841	0.989328
3	40.0	0.106869	0.219096	0.982605
4	50.0	0.062652	0.188462	0.989610
5	60.0	0.063964	0.190277	0.989603
6	70.0	0.064061	0.192848	0.989426
7	80.0	0.108336	0.220430	0.982368
8	90.0	0.063216	0.187610	0.989575
9	100.0	0.059805	0.188470	0.990059

Figure 7 Performance metrics corresponding to the best performing models.

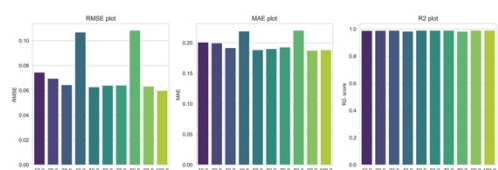


Figure 9 Results of hyperparameter tuning in RF.

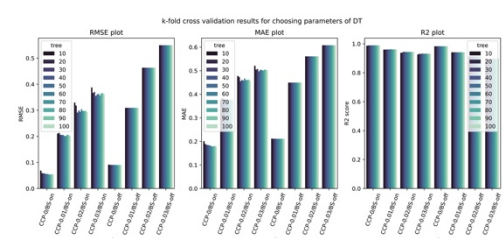
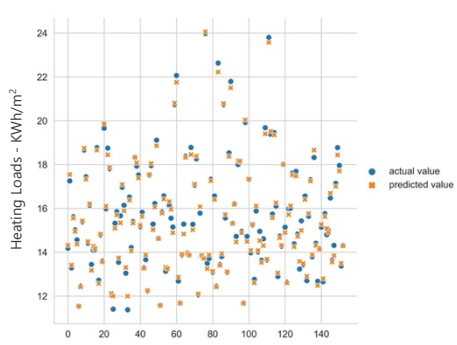


Figure 8 Actual versus predicted values for randomly selected test cases.



Experiment (3): Cell complex (cooling loads):

The result of k-fold cross validation is given in Figure 9. The results of this experiments revealed the following:

- In general, the performance showed improvement as the number of trees increased, except in the cases of choosing 40 or 80 trees, which led to a decrease in performance.
- The performance was significantly worse when the bootstrap was disabled in comparison to when it was active.
- According to the experiments, the optimal model consisted of a bootstrap that was enabled, a ccp-alpha value of 0, and 100 trees Table 3.

Figure presents the performance metrics of the top-performing models based on the number of trees. The final test metrics were: RMSE of 0.053527, MAE of 0.179106, and an R2 score of 0.989864. A visualization of RF predictions for a randomly energy data point is depicted in Figure 11.

	Tree	RMSE	MAE	R^2- score
0	10.0	0.068589	0.200871	0.986969
1	20.0	0.058965	0.189458	0.988896
2	30.0	0.058229	0.184866	0.988994
3	40.0	0.090525	0.211556	0.983036
4	50.0	0.057170	0.184962	0.989190
5	60.0	0.055689	0.182632	0.989460
6	70.0	0.055686	0.182269	0.989446
7	80.0	0.090443	0.211525	0.983044
8	90.0	0.054071	0.178023	0.989763
9	100.0	0.053527	0.179106	0.989864

Table 3 K-fold cross validation for RF.

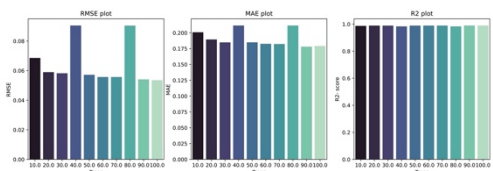


Figure 10 Performance metrics corresponding to the best performing models.

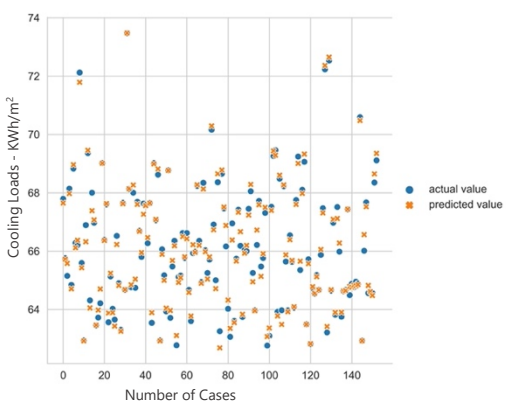


Figure 11 Actual versus predicted values for randomly selected test cases.

Experiment (4): Cell complex (heating loads):

The k-fold cross-validation outcomes are displayed in Figure 12. Insights from these experiments include:

- Performance demonstrated an oscillatory trend with the addition of more trees, with optimal results observed at 60 trees.
- Consistent with experiment (3), performance declined when the bootstrap option was turned off.
- Based on the experiments, the best model configuration featured an enabled bootstrap and 60 trees.

Figure 13, shows the performance metrics of the best-performing models, sorted according to the number of trees used. The final test metrics were: RMSE of 0.060729, MAE of 0.190074, and an R2 score of 0.987454. A visualization of RF predictions for a randomly energy data points is depicted in Figure 14.

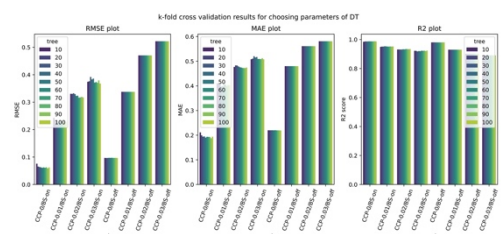


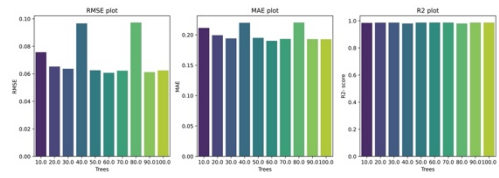
Figure 12 The results of hyperparameter tuning in RF.

	Tree	RMSE	MAE	R^2- score
0	10.0	0.075701	0.211335	0.984441
1	20.0	0.065312	0.199333	0.986448
2	30.0	0.063677	0.194189	0.986750
3	40.0	0.096643	0.219836	0.980213
4	50.0	0.062563	0.195042	0.987012
5	60.0	0.060729	0.190074	0.987454
6	70.0	0.062288	0.193283	0.987083

Table 4 The results of k-fold cross validation for RF.

7	80.0	0.097308	0.220311	0.980056
8	90.0	0.061227	0.193029	0.987271
9	100.0	0.062436	0.192772	0.987068

Figure 13 Performance metrics corresponding to the best performing models.



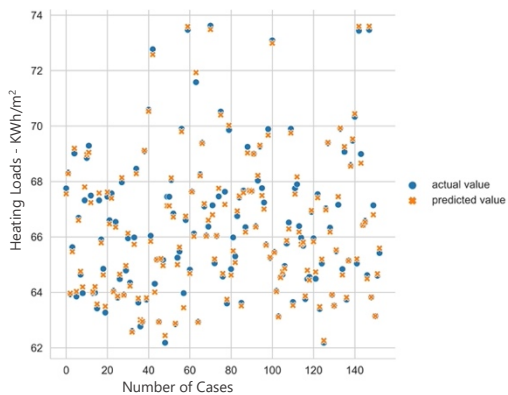


Figure 14 Graph shows actual versus predicted values for randomly selected test cases.

CONCLUSION

This study introduced an innovative method that utilises machine learning (ML) algorithms to predict cooling and heating loads in the early stages of building design. While not empirically tested within the paper, the inherent characteristics of surrogate models like RF suggest potential time savings. Traditional simulation methods involve setting up detailed models and running computationally expensive simulations, which can be time-consuming and disrupt the creative process of architects. In contrast, once trained, the RF model can provide rapid predictions, facilitating quicker decision-making in the early design stages.

In general, the results highlighted the considerable capacity of surrogate modelling to precisely predict energy loads in buildings. The Random Forest (RF) model was trained using a parametric generative methodology, which involved creating an energy dataset using TopologicPy for model development. The investigation included two scenarios: one using a single building unit referred to as a 'cell,' and another including complex architectural arrangements known as 'cell complexes.' Both proposals were simulated to predict the cooling and heating loads of the building. Afterwards, the energy data produced by the simulations were used to create and verify the RF model. Four experiments were conducted to analyse the cooling and heating loads by investigating both single cell and cell complex structures. In order to select the best model, each experiment employed k-fold cross-validation to split the data, along with hyperparameter modifications.

The study revealed that the k-fold cross-validation procedure improved the performance of the model by ensuring that each fold was used for both training and testing, hence reducing bias caused by binary splits. Furthermore, making alterations to the hyperparameters had a substantial impact on the accuracy of the RF model. The results indicated that the RF model was able to accurately predict cooling and heating demands with high precision as follows: Experiment 1 (R2: 0.991776), Experiment 2 (R2: 0.990059), Experiment 3 (R2: 0.989864), and Experiment 4 (R2: 0.987454). These results demonstrate a high level of predicted accuracy which are consistent with other performance metrics. The results indicate that RF models show potential for accurately predicting cooling and heating loads in buildings during the first design phases.

We recognize the importance of empirically validating the time-saving benefits of the proposed approach. Future research will focus on conducting comparative studies to quantify the time savings provided by RF models compared to traditional simulation techniques. This will involve direct measurements of setup and computation times for both methods across various case

studies. Due to the timeframe of this research and the computation time, this study examined a specific range of parameters to train the RF models, limiting the generalizability of the developed surrogate model to predict other scenarios. We acknowledge the limitations and constraints of generalizing surrogate models to different scenarios. Therefore, future work will focus on generating a wider variety of parameters to enhance the model's generalizability to different design settings.

Performance metric	Experiment 1	Experiment 2	Experiment 3	Experiment 4
RMSE	0.050551	0.059805	0.053527	0.060729
MAE	0.171947	0.188470	0.179106	0.190074
R2-score	0.991776	0.990059	0.989864	0.987454

Table 5 The performance metrics for all experiments.

REFERENCE

Ahmad, M.W. et al. 2017. Random Forests and Artificial Neural Network for Predicting Daylight Illuminance and Energy Consumption. *5th Conference of International Building Performance Simulation Association* , pp. 1–7.

Alammar, A. et al. 2021. Predicting Incident Solar Radiation on Building ' s Envelope Using Machine Learning.

Alammar, A. and Jabi, W. 2023. *Generation of a Large Synthetic Database of Office Tower's Energy Demand Using Simulation and Machine Learning* . Springer Nature Singapore. Available at: http://dx.doi.org/10.1007/978-981-99-2217-8_27.

Chou, J.S. and Bui, D.K. 2014. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings* 82(2014), pp. 437–446. Available at: http://dx.doi.org/10.1016/j.enbuild.2014.07.036.

Jabi, W. et al. 2018. Topologic A toolkit for spatial and topological modelling. *Proceedings of the International Conference on Education and Research in Computer Aided Architectural Design in Europe* 2(September), pp. 449–458. doi: 10.52842/conf.ecaade.2018.2.449.

Jabi, W. 2022. Topologic and BHoM : Enhancing Energy Analysis Workflows Through Topological Modelling. (October 201

Liu, Y. et al. 2021. Enhancing building energy efficiency using a random forest model: A hybrid prediction approach. *Energy Reports* 7, pp. 5003–5012. Available at: https://doi.org/10.1016/j.egyr.2021.07.135.

Rashidifar, R. 2020. Estimation of {Energy} {Performance} of {Buildings} {Using} {Machine} {Learning} {Tools}. (Cl). Available at: http://dx.doi.org/10.31224/osf.io/4esdw.

Sarmas, E. et al. 2023. Estimating the Energy Savings of Energy Efficiency Actions with Ensemble Machine Learning Models. *Applied Sciences (Switzerland)* 13(4). doi: 10.3390/app13042749.

Seyedzadeh, S. et al. 2018. Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering* 6(1). doi: 10.1186/s40327-018-0064-7.

Tso, G.K.F. and Yau, K.K.W. 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* 32(9), pp. 1761–1768. doi: 10.1016/j.energy.2006.11.010.

Westermann, P. and Evins, R. 2019. Surrogate modelling for sustainable building design – A review. *Energy and Buildings* 198, pp. 170–186. Available at: https://doi.org/10.1016/j.enbuild.2019.05.057.

Yu, Z. et al. 2010. A decision tree method for building energy demand modeling. *Energy and Buildings* 42(10), pp. 1637–1646. Available at: http://dx.doi.org/10.1016/j.enbuild.2010.04.006.