# Comparison of three random forest models of a chiller system

Han Sol Shin[1], Cheol Soo Park[1*]
[1]School of Civil, Architectural Engineering and Landscape Architecture,
Sungkyunkwan University, Suwon, South Korea
[*]Corresponding author: cheolspark@skku.ac.kr

## Abstract

The current building operation and maintenance is dependent on subjective decisions, e.g. building operator's experience and knowledge, rather than employing a simulation model-assisted operation. It demands in-depth knowledge of building physics, systems and controls to develop the simulation model for optimal operation. Rather than using the first-principles based simulation tools, this paper presents a machine learning simulation model of a chiller in an office building. For this study, the BEMS data (a chiller's electric energy, chiller supply water temperature, AHU return water temperature, AHU water flow rate, etc.) were collected from the existing office building (a total floor area: 21,577m$^2$). The authors used a Random Forest (RF) method, one of the machine learning techniques. Three RF models were developed and cross-compared in this study as follows: Model A developed with 12 variables from BEMS data, Model B developed with the 12 variables plus 18 new variables constructed by two arithmetic operators (a total of 20 variables), Model C with the 12 variables plus 6 new variables constructed based on physics-based equations (a total of 18 variables). The CVRMSE of the three models are 8.56%, 5.44% and 4.28%, respectively.

## Introduction

Energy consumed in buildings takes up more than 40% of national energy consumption (IEA, 2013), and half of the energy is consumed in buildings can be saved through energy-efficient design and optimal control of building systems (Baird, 1984). However, the current building control is generally based on the subjective judgment, experience and knowledge of building operators, rather than using a dynamic simulation model.

It requires significant time and effort to develop an accurate dynamic simulation model of building systems. In addition, for the development of such simulation model, there are many uncertain inputs, such as convective/radiative heat transfer, air movement in and around the building, etc. In the development process, many subjective assumptions and simplifications of the reality are also involved. This causes the issues such as transparency, reproducibility, and objectivity of the simulation model and its use for control and maintenance (Ahn, 2015).

A data-driven machine learning model can be effectively used to imitate the dynamic behavior of building systems. Compared to the first-principles based simulation model, the machine learning model can be developed with fewer inputs and less time and effort (Kim, 2016). In this study, the authors developed a simulation model of a chiller (1,250 kW, nominal COP 5.53) in an office building(5 floors above ground, 21,577 m$^2$). The building's operational data (the chiller's supply/return water temperature, cold water flow rate, chiller's electric energy use, etc.) were collected during August. The chiller model was constructed by random forest (RF) algorithm. Three different RF models (Model A, B, C) were developed to investigate the degree of model accuracy depending on the degree of expertise and knowledge.

## Random forest

Random forest algorithm has been used in various fields for classification, clustering and regression: object tracking (Schulter et al, 2014; Gall et al, 2009), picture classification (Ristin et al, 2014; Bosch et al, 2007), and corporate credit risk management (Brown et al, 2012; Kalsyte et al, 2013), etc. The RF algorithm is a kind of ensemble method proposed by Breiman (2001), combining the random input selection (Amit et al, 1997) with the bagging (Bootstrap Aggregating). The ensemble method, a kind of machine learning methods, is used to predict the state of a system by combining several decision trees. The method calculates a final result by the average of the predictions of each decision tree in the case of regression, and by voting in the case of classification.

The development process of the RF model is shown in Figure 1 and the process is as follows:

(1) Bootstrap sampling: From training data, *n* bootstrap samples are randomly sampled with replacement (Louppe, 2014). Each sample is used as training data to build a decision tree.

(2) Decision tree growth: The best binary split is found among *m* input variables which is randomly selected. For each bootstrap sample, the pruning is not performed when the tree is fully grown..

(3) Ensemble: Ensemble *n* decision trees and construct them as one random forest model. When new data are used to the model, each decision tree calculate the prediction and random forest output is an average of all predictions.

(4) OOB MSE: Out-of-bags Mean Squared Error (OOB MSE) is generated from OOB samples (Equation 1). $N_{t,OOB}$ is a data size of the t[th] decision tree, $Y_{t,OOB}$ is an observed value of the t[th] decision tree, $\bar{Y}_{t,OOB}$ is a mean of a predicted value of the t[th] decision tree. About 37% of the training data are not included in the bootstrap sample and extracted as OOB sample. The user can verify whether the appropriate number of decision trees are generated or not with the convergence of OOB error rate.

$$\text{OOBMSE} = \frac{1}{N_{t,OOB}} \sum_{t=1}^{N_{t,OOB}} \left(Y_{t,OOB} - \bar{Y}_{t,OOB}\right)^2 \qquad (1)$$

(5) Variable Importance (VI): VI is an index that measures whether the input variable is used as an important classifier in the model, using the permuted OOB data (Equation 2). $Y_{j,p}$ is a predicted value of the j[th] input variable, $Y_j$ is a predicted value of the j[th] input variable, $\sigma$ is a standard deviation. VI measures how well each input can predict the output.

$$\text{VI} = \frac{OOBMSE\left(Y_{j,p}\right) - OOBMSE\left(Y_j\right)}{\sigma\left(OOBMSE\left(Y_{j,p}\right)\right) - \sigma\left(OOBMSE\left(Y_j\right)\right)} \qquad (2)$$

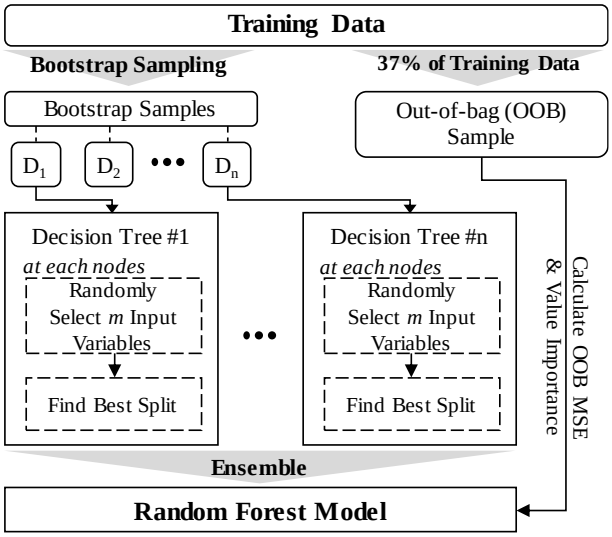
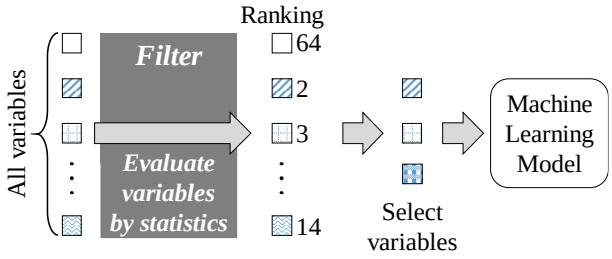



*Figure 1: Modeling process of random forest algorithm*

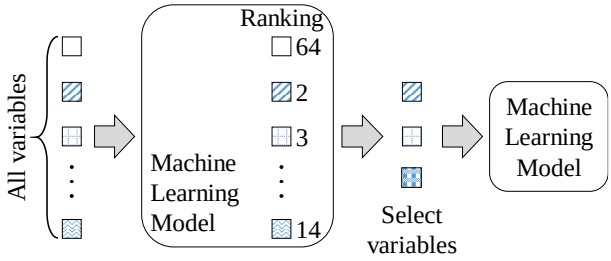
## Variable construction and selection

Variable construction is necessary to generate input variables having a significant correlation with output variables. The generated input variables do not always have a physical meaning (Kotsiantis et al, 2006). It is common to randomly generate new input variables out of raw input variables or measured data from BEMS.

After variable construcion is done, the process of 'variable selection' is required for RF modeling. Variable selection reduces the size of input variables as well as computation time by removing unnecessary variables. Two typical methods of variable selection are as follows (Guyon et al, 2003; May et al, 2011): (1) filter and (2) wrapper. The 'filter' method selects input variables independently from the model (Figure 2(a)), while the 'wrapper' method, as inferred by its name, selects input variables inside the model (Figure 2(b)). The latter method takes more computation time, but provides better performance in variable selection. In this paper, the authors used the 'wrapper' method.




*(a) Filter*




*(b) Wrapper*

*Figure 2: Variable selection*

## Target building and chiller system

The target building, a total floor area of 21,577 m$^2$, is an office building located in Seoul, Korea. A Building Energy Management System (BEMS) is installed in the building and store relevant operation information in real time. The stored data include the chiller's electric energy use, water supply / return temperature, air handling unit (AHU) / fan coil unit (FCU) header water flow, etc. The chiller in the target building is a compression-type chiller and its rated capacity is 1,250 kW and nominal COP is 5.53 (Figure 3).

*Figure 3: Chiller system in the target building*

For this study, the aforementioned data were measured for 5 days from 00:00 on August 12th to 24:00 on August 16th at a sampling time of 5 minutes. The data measured for the first three days (from 12th to 14th) and for the last two days (from 15th to 16th) were used as the training

data and the validation data, respectively. The data measured for the five days are shown in Table 1.

*Table 1: The BEMS data*

| Name | Measured variables | | Unit |
|---|---|---|---|
| y | Chiller | Electric energy use | kW |
| x1 | | Running state | on/off |
| x2 | | Supply water temperature | ℃ |
| x3 | | Return water temperature | ℃ |
| x4 | AHU header | Return water temperature | ℃ |
| x5 | | Supply water temperature | ℃ |
| x6 | | Water flow rate | kg/h |
| x7 | FCU header | Supply water temperature | ℃ |
| x8 | | Return water temperature | ℃ |
| x9 | | Water flow rate | kg/h |
| x10 | Cooling tower | Supply water temperature | ℃ |
| x11 | | Return water temperature | ℃ |
| x12 | Weather | Outdoor air temperature | ℃ |

## Model development

The authors used a MATLAB toolbox "M5' regression tree" made by M5PrimeLab (Jekabsons, 2016). In the random forest modeling, the following parameters were to be determined: the number of trees ($n$) and the number of input variables ($m$) at each node. In this study, the authors used 300 trees and the number of input variables was set at $N/3$, recommended by (Hastie et al, 2008). $N$ means the number of the entire input variables.

### Model A: 12 input variables obtained from BEMS

In Model A, 12 input variables (x1-x12, Table 1) obtained from BEMS were used. As shown in Figure 4 as the number of trees increases, the accuracy of the chiller model improves. After the number of trees is close to 300, the OOB MSE converges to 50 kW. It means that 300 trees are good enough for Model A. Figure 5 shows the comparison between model prediction and measured data based on the training set (August 12th to 14th). The model's Mean Bias Error (MBE), Coefficient of Variation of Root Mean Squared Error (CVRMSE) and Root Mean Squared Error (RMSE) are 0.02%, 4.61% and 3.79kW, respectively. Figure 6 shows a part of random forest model of the chiller's electric energy use. For example, if input variable x7 exceeds 19.5 and x10 exceeds 31.1, the predicted electric energy use of the chiller is 437kW (the red line in Figure 6).

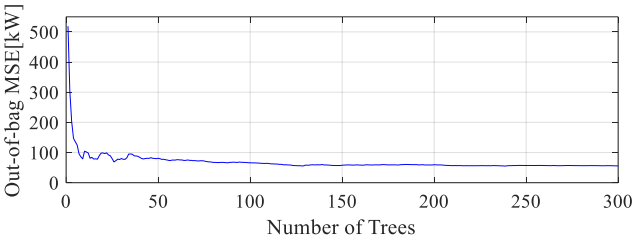



*Figure 4: OOB MSE of Model A*

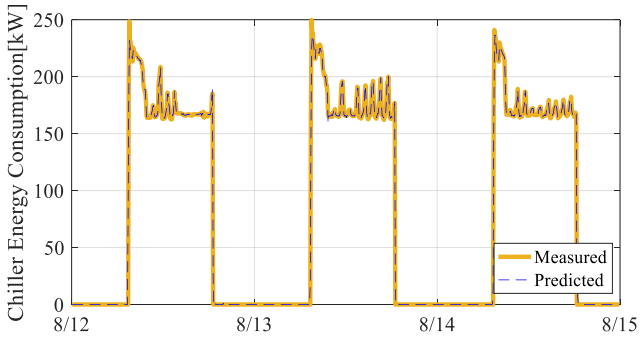



*Figure 5: Comparison between model prediction and measured data (training set, Aug. 12-14)*

### Model B: 12 plus 18 newly constructed variables

Additional 18 variables were constructed. This variable construction was performed by randomly choosing 3 variables and 2 arithmetic operations. The 18 input variables were added to the existing 12 input variables (Table 2).

*Table 2: Variable construction (Model B)*

| Name | Combinations | Name | Combinations |
|---|---|---|---|
| p1 | x5 × x6 − x4 | p10 | x8x × 9x − 11 |
| **p2** | x6 ÷ x3 − x3 | p11 | x3 × x3 × x8 |
| p3 | x4 − x2 + x6 | p12 | x2 − x4 ÷ x3 |
| p4 | x8 − x1 × x2 | p13 | x10 − x1 + x3 |
| p5 | x6 × x13 − x5 | p14 | x10 + x4 × x3 |
| p6 | x11 + x9 ÷ x6 | **p15** | x8 + x11 ÷ x1 |
| p7 | x12 × x8 + x8 | **p16** | x6 − x12 − x2 |
| p8 | x8 ÷ x7 × x11 | p17 | x12 + x5 ÷ x2 |
| p9 | x11 + x8 × x10 | **p18** | x7 × x10 + x5 |

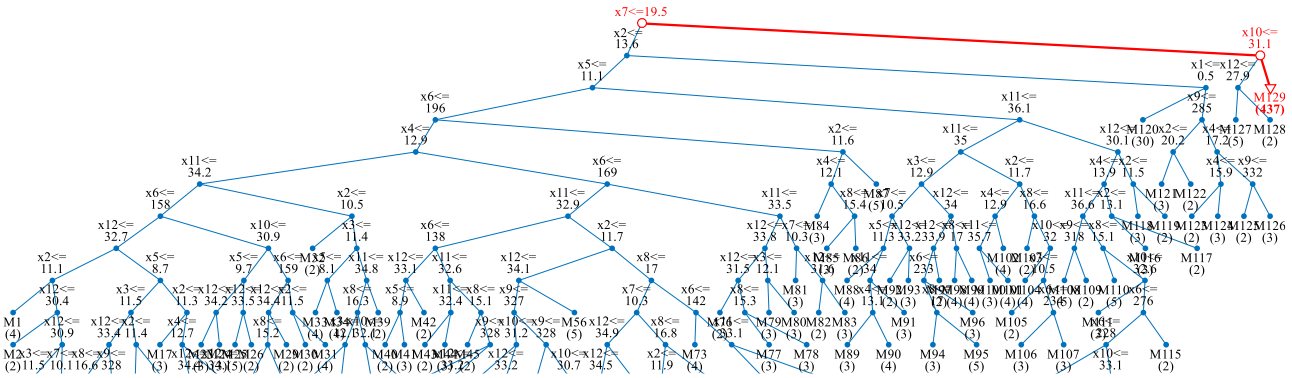



*Figure 6: A part of random forest model of Model A*

Figure 7, calculated from RF method, shows variable importance of a total 30 input variables. The newly generated variable, p2, p15, p18, p16 are found to be the most correlated with the chiller's electric energy use. This means that the simple arithmetic relation can improve the correlation between input variables and output variables.

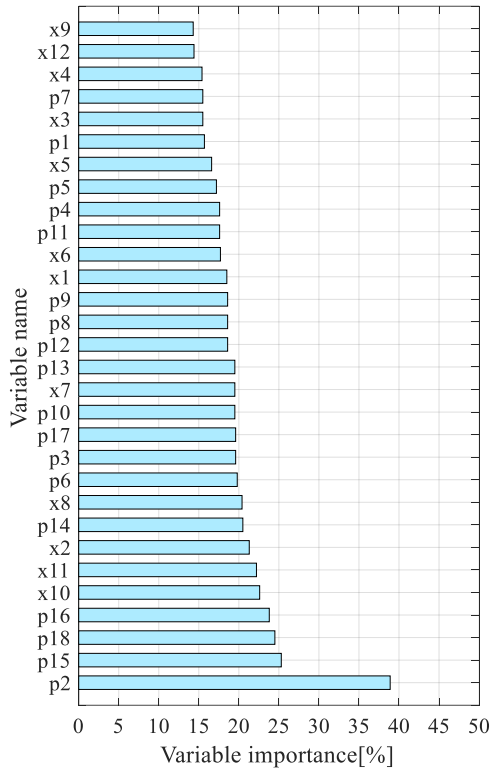



*Figure 7: Variable importance (Model B)*

In RF modeling, the authors applied the wrapper method for variable selection as mentioned in the earlier section. In the wrapper method, a backward variable elimination was used. As shown in Figure 8, the MBE, RMSE, and CVRMSE begin to increase when the number of eliminated input variables becomes greater than 25. The authors selected six input variables (x10, x11, p2, p15, p16, p18) for Model B. Thus, the number of eliminated input variables is 24. The CVRMSE of Model B is 5.44%.

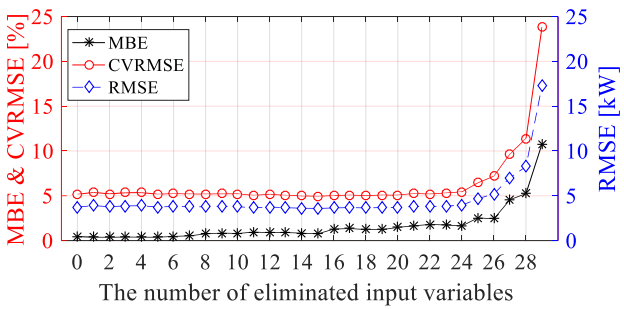



*Figure 8: Accuracy (Model B)*

**Model C: 12 variables plus 6 new variables constructed based on physics-based equations**

In the process of developing Model C, the authors applied physics-based equations to variable construction. The chiller's electric energy use is correlated with the amount of heat removed from the chiller. Thus, new input variables were added as shown in Table 3. The total number of input variables are 18. The input variables of the significant variable importance are the difference between supply and return water temperature at the chiller (q1), and the difference between supply/return water temperature at the cooling tower (q6) (Figure 9).

*Table 3: Variable construction (Model C)*

| Name | Constructed variables | | Unit |
|---|---|---|---|
| **q1** | Chiller | $T_{out} - T_{in}$ | ℃ |
| q2 | AHU header | $T_{out} - T_{in}$ | ℃ |
| q3 | FCU header | $T_{out} - T_{in}$ | ℃ |
| q4 | AHU header | $C \times \dot{m} \times (T_{out} - T_{in})$ | kW |
| q5 | FCU header | $C \times \dot{m} \times (T_{out} - T_{in})$ | kW |
| **q6** | Cooling tower | $T_{out} - T_{in}$ | ℃ |

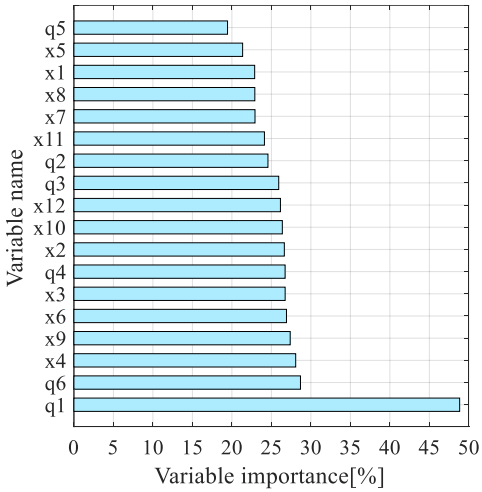



*Figure 9: Variable importance (Model C)*

Figure 10 shows the result of the wrapper method for Model C. The MBE, RMSE and CVRMSE of Model C rises begin to increase when the number of eliminated input variables becomes 13. Model C was finally developed with 6 input variables (x3, x4, x6, x9, q1, q6) and the resulting CVRMSE is 4.28%.

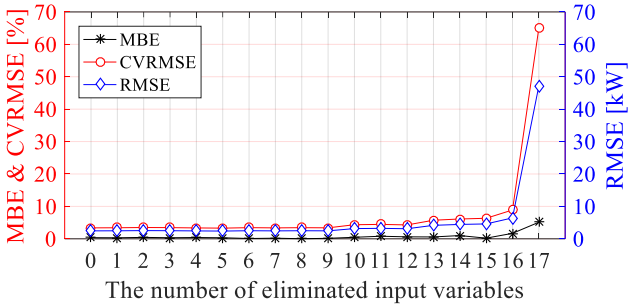
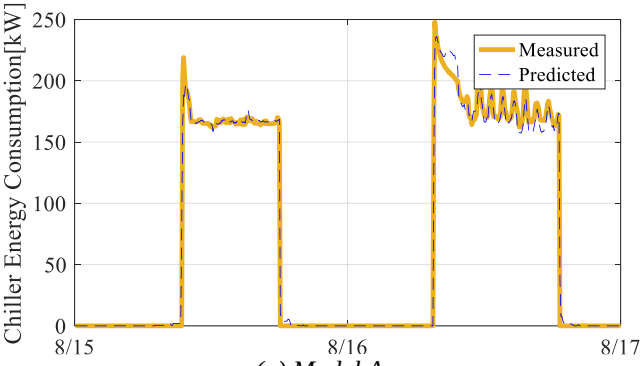



*Figure 10: Accuracy (Model C)*

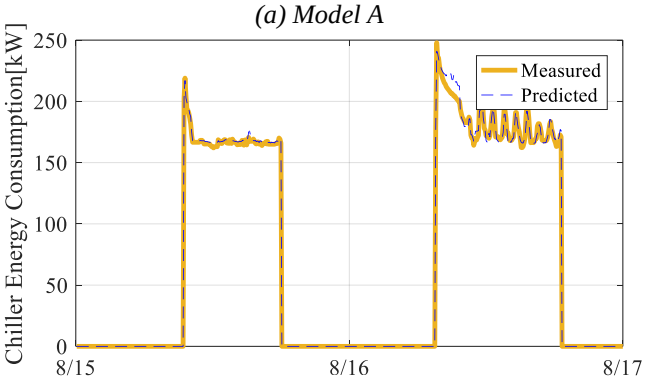
**Validation and discussion**

For the validation purpose, the authors used the measured data for the last two days out of the five days' experiment. As shown in Figure 11 and Table 4, all three models are good enough for predicting the chiller's energy use. Model B and C use far less input variables than Model A but both can perform satisfactorily. The newly generated input variables for Model B (p2, p15, p18, p16) and for

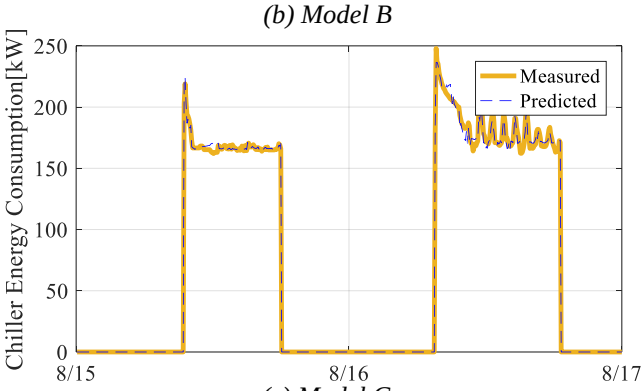
Model C (q1, q6) proved to be top 4 and top 2 variables for Model B and C in terms of VI (Table 5). Contrary to the author's expectation, there is no significant difference in MBE, CVRMSE, and RMSE between Models A, B and C.




*(a) Model A*




*(b) Model B*




*(c) Model C*

*Figure 11: Comparison between model prediction and measured data (Aug. 15-16)*

*Table 4: Comparison of three models*

| Model | The number of input variables | MBE | CVRMSE | RMSE |
|---|---|---|---|---|
| A | 12 | 0.14 % | 8.56 % | 6.21 kW |
| B | 6 | 1.64 % | 5.44 % | 3.95 kW |
| C | 6 | 0.60 % | 4.28 % | 3.11 kW |

*Table 5: Variable importance ranking comparison between three models*

| Ranking | Model A | Model B | Model C |
|---|---|---|---|
| 1 | x1 | p2 | q1 |
| 2 | x6 | p15 | q6 |
| 3 | x11 | p18 | x4 |
| 4 | x7 | p16 | x9 |
| 5 | x4 | x10 | x6 |
| 6 | x2 | x11 | x3 |

## Conclusion

For optimal control of the chiller, a simulation model was developed by RF algorithm. The authors developed three different models based on the assumption that as more physical input variables are included in the model, the model will become more accurate.

It is shown in the paper that three models can predict the chiller system accurately. Contrary to our expectation, the two RF models (Model B and C) with new generated input variables perform marginally better than the model constructed with raw data (Model A). It can be inferred that the machine learning algorithm itself is good enough to generate a simulation model, without requiring any additional expertise and in-depth prior knowledge, as indicated in the accuracy of Model A.

As a future study, application of machine learning models to Model Predictive Control (MPC) will be sought.

## Acknowledgement

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Trade, Industry & Energy(MOTIE) of the Republic of Korea (No. 20152020105550).

## References

Ahn, K.U., Kim, D.W., Kim, Y.J., Park, C.S., and Kim, I.H. (2015). Gaussian Process Model for Control of an Existing Building. *Energy Procedia*, *78*, 2136-2141.

Amit, Y., and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, *9*(7), 1545-1588.

Baird, G., Aun, C., Brauder, W., Donn, M . R., and Pool, F. (1984). Energy performance of buildings.

Bosch, A., Zisserman, A., and Munoz, X. (2007). *Image classification using random forests and ferns.* Paper presented at the 2007 IEEE 11th International Conference on Computer Vision.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brown, I., and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453.

Gall, J., and Lempitsky, V. (2013). Class-specific hough forests for object detection *Decision forests for computer vision and medical image analysis* (pp. 143-157): Springer.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

Huang, J., Li, Y. F., and Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology*, *67*, 108-127.

Iliou, T., Anagnostopoulos, C. N., Nerantzaki, M., and Anastassopoulos, G. (2015). *A Novel Machine*

*Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance.* Paper presented at the Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS).

Jekabsons, G. M5'regression tree, model tree, and tree ensemble toolbox for Matlab/Octave ver. 1.6. 0.

Kalsyte, Z., and Verikas, A. (2013). A novel approach to exploring company's financial soundness: Investor's perspective. *Expert Systems with Applications, 40*(13), 5085-5092.

Kim, Y. M., Ahn, K. U., and Park, C. S. (2016). Issues of Application of Machine Learning Models for Virtual and Real-Life Buildings. *Sustainability*, *8*(6), 543.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science, 1*(2), 111-117.

Kusiak, A., Li, M., and Zhang, Z. (2010). A data-driven approach for steam load prediction in buildings. *Applied Energy, 87*(3), 925-933.

Louppe, G. (2014). *Understanding random forests: From theory to practice.* (Ph.D.), University of Liège, Liège.

May, R., Dandy, G., and Maier, H. (2011). *Review of input variable selection methods for artificial neural networks*: INTECH Open Access Publisher.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). Machine learning, neural and statistical classification.

Platon, R., Dehkordi, V. R., and Martel, J. (2015). Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. *Energy and Buildings, 92*, 10-18.

Ristin, M., Guillaumin, M., Gall, J., and Van Gool, L. (2014). *Incremental learning of NCM forests for large-scale image classification.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Schulter, S., Leistner, C., Wohlhart, P., Roth, P. M., and Bischof, H. (2014). *Accurate object detection with joint classification-regression random forests.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Tooke, T. R., Coops, N. C., and Webster, J. (2014). Predicting building ages from LiDAR data with random forests for building energy modeling. *Energy and Buildings, 68*, 603-610.

Trevor, H., Robert, T., and Jerome, F. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2 ed.): Springer-Verlag New York.

Tsanas, A., and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings, 49*, 560-567.

Yan, K., Shen, W., Mulumba, T., and Afshari, A. (2014). ARX model based fault detection and diagnosis for chillers using support vector machines. *Energy and Buildings, 81*, 287-295.

Yang, J., Santamouris, M., Lee, S. E., and Deb, C. (2016). Energy performance model development and occupancy number identification of institutional buildings. *Energy and Buildings, 123*, 192-204.