Improving Prediction Accuracy of Machine Learning Energy Prediction Models

Manav Mahan Singh[1*], Sundaravelpandian Singaravel[1] and Philipp Geyer[1]
[1]Architectural Engineering Division, Department of Architecture, KU Leuven
* email: manavmahan.singh@kuleuven.be

Abstract

Machine learning (ML) energy prediction models are useful to estimate the building energy demand with short response time, particularly important for developing energy-efficient building designs at an early stage of design. Component-based machine learning model (CBML) is the state-of-the-art development of ML energy prediction and useful to improve the generalizability of models. The concept of CBML is based on decomposing the design artefact in an engineering-related way and calculating intermediate parameters such as heat flows through building elements, followed by energy component at zone level and finally total energy demand at building level. Previous research showed that the method works but the improvement of the range of design and accuracy are still desirable to make the method more aligned to architectural design. Therefore, in this paper, we propose to enrich the training data with new building shapes and enhanced features typical for architectural design. For training each model of CBML, the data is collected by performing parametric energy simulations with three different building shapes and tested on fourth shape. A manual feature engineering approach is carried out by extracting useful features which have an influence on the target value. The accuracy of CBML is ascertained by *coefficient-of-determination (R^2-values)* and *mean absolute percentage error (MAPE)*. The effect of enriching data with different building shape is studied by training ML model while increasing data sequentially and recording the improvements in the prediction accuracy. To study the effect of enhancing feature, CBML is developed with two types of features – raw and enhanced features and recording the change in the prediction accuracy. Furthermore, the influence of features is calculated using permutation importance to study the effect of additional features. The accuracy of total building energy model on new building shape improves from 5.18% to 3.14% *(MAPE)* and 0.9970 to 0.9988 *(R^2)* after enriching the data with several building shapes and enhancing the features.

Keywords: Feature Engineering, Training Data, Feature Importance, Performance, Permutation Importance.

1. Introduction

Machine learning (ML) energy prediction models are useful to predict the building's energy demand in a short response time compare to traditional energy simulation tool (Geyer & Singaravel, 2018; Singaravel, Geyer, & Suykens, 2018). This is quite relevant for predicting energy demand and developing energy-efficient building designs at an early stage of design when the design parameters are inherently uncertain (Struck, de Wilde, Hopfe, & Hensen, 2009; Tian et al., 2018; Van Gelder, Janssen, & Roels, 2014). The concept of component-based machine learning model (CBML) is introduced by Geyer and Singaravel to improve the ML model generalization to design cases not present within training design cases (Geyer & Singaravel, 2018). In monolithic models, target variable is directly predicted using building design parameters as input features. However, the concept of CBML is based on decomposing the design artefact in an engineering-related way and calculating intermediate parameters. CBML offers a unique opportunity to integrate an energy prediction model easily with multi-level-of-detail (multi-LOD) building information model (BIM) data structure (Abualdenien & Borrmann, 2018; Geyer, Singh, & Singaravel, 2018). The integration of CBML with BIM models has the potential to streamline the design and energy prediction process at any stage of the design process (Singh, Singaravel, & Geyer, 2018).

In previous implementations of the CBML, a decrease of the accuracy is observed when using component models trained on parametric simulation results from rectangular buildings on more complex building shapes (Geyer & Singaravel, 2018). Reason being complex environmental interactions present for complex building shapes were not present within the training space characterized by rectangular buildings. In this research work, the accuracy of CBML is improved by enriching the training data with several building shapes and enhancing features. It has been proven that the accuracy of ML models increases with the training data irrespective of the algorithm in speech learning application (Banko & Brill, 2001). The approach of enhancing features to improve model accuracy is discussed in few studies (Catalina, Virgone, & Blanco, 2008; Cheng & Cao, 2014). The accuracy of the model improves with newly developed features, but it lacks the use of formal techniques to study the feature importance. The objectives of this study are:

1) To study the effect of enriching data with several building shapes to improve the prediction accuracy of ML models.
2) To study the effect of enhancing features on the prediction accuracy of ML models.

2. Literature Review

There are several efforts made to make quick energy prediction using engineering-surrogate or data-driven models (Van Gelder, Das, Janssen, & Roels, 2014). These models offer few advantages over traditional energy simulation tools in terms of computational time at the cost of prediction accuracy. It is possible to evaluate a large number of simulation model using these approaches which are required for making probabilistic energy prediction at the early stage of design (Van Gelder, Janssen, et al., 2014). There are several research studies published on the use of machine learning models for energy prediction (Fumo, 2014). However, these models offer limited integration with the design process and the applicability of the models is limited to the training design case. The concept of CBML is introduced to overcome multiple limitations of monolithic ML energy prediction models such as extensionality to new design cases, training models for generic building elements, and integration with BIM model (Geyer & Singaravel, 2018).

In other machine learning applications such as speech recognition, it has been proven that enriching training data with the new cases improves the prediction accuracy of the models (Banko & Brill, 2001; Halevy, Norvig, & Pereira, 2009). Van Gelder, Das, et al., 2014 studies the effect of increasing the number of samples on the prediction accuracy, but the effect of enriching training data with various building shapes is not tested on the energy prediction models. There are few studies which utilize data from various building shapes for the training of energy prediction models, but the prediction accuracy is never tested on the building shape outside of the training data (Asadi, Amiri, & Mottahedi, 2014; Catalina et al., 2008). A manual feature engineering approach was carried out by using domain knowledge and extracting useful features which have an influence on the target value (Catalina et al., 2008). Previously developed machine learning energy prediction approaches document limited applicability of the models on new design cases or the accuracy of the models reduces.

3. Structure of CBML and training of ML model

The concept of CBML is based on decomposing the design artefact in an engineering-related way and calculating intermediate parameters. In the presented approach the building is composed of a zone which has building elements such as walls, windows, floors and roofs. Figure 1 shows the CBML architecture utilized in this paper. ML models at building elements like walls first predict heat flows through them. Information on heat flows together with design information like area are utilized by ML models at zone level to make predictions on zone energy demand. Finally, ML model at building level utilizes information of zone energy demand together with HVAC design information to make the final energy prediction. Five types of building elements are considered in this model, namely, *Wall, Window, Ground Floor (GFLoor), Intermediate Floor (IFloor)* and *Roof*. The heat flows through the building components are summed up as *Total Heat Flows* and provided as input for zone level models i.e. *Heating Energy, Cooling Energy, Lighting Energy* and *Equipment Energy*. The zone level outputs are used to train the *Building Total Energy* model. The complete list of features used for training of each model of CBML is available *Table 2*.

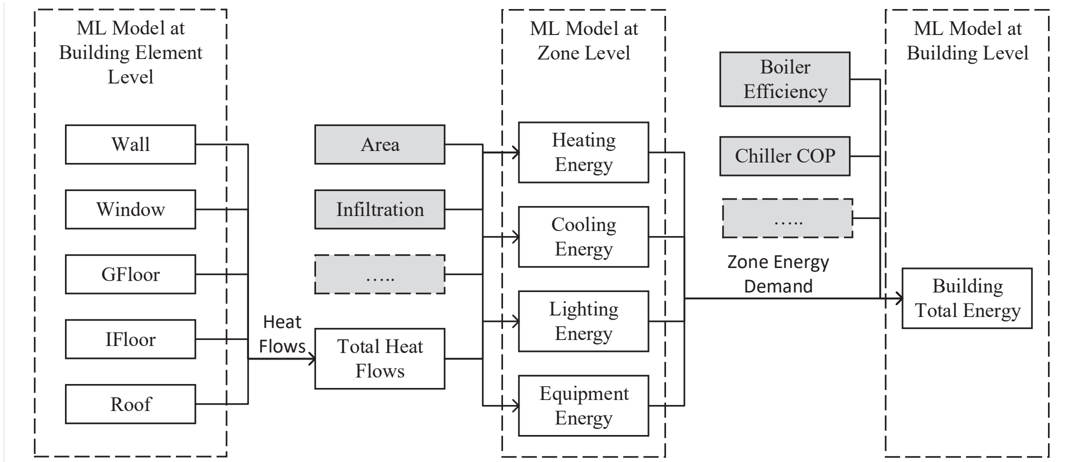

Figure 1. Structure of component-based machine learning model

The models have two hidden layers with a specified number of neurons. There are three hyper-parameters which are the number of neurons in first and second hidden layer (NN1 and NN2) and L2-regularisation which are tuned to obtain a suitable model. L2-regularisation is used to avoid overfitting issue by adding a squared magnitude of coefficients as penalty term to the loss function (Ng, 2004). The few combinations of hidden layers NN1 {10, 20}, NN2 {4, 8} and L2-regularisation {0.001, 0.0001, 0.00001} are tested to identify a good combination of hyper-parameters for each model. These hyper-parameters are tuned each time using the validation data which is 20% of training data. The performance of each model of CBML is estimated by two parameters which are coefficient of determination (R^2) and mean absolute percentage error (MAPE). R^2 represents a measure of second order and it is more sensitive to mid-range values while MAPE is a measure of the first order and more suitable to low-range values (Géron, 2017). Thus, both parameters are important to assess the model accuracy. ML models are developed using Keras and TensorFlow as a backend (Chollet, 2015).

4. Research methodology

This section consists of three sub-sections. The section 4.1 details out the data collection process which is used to train ML models. The section 4.2 explains the method for studying the effect of enriching the training data. The section 4.3 elaborates the details the method to study the effect of enhancing features.

4.1 Data collection for training ML models

We have collected data using parametric simulation with four different building shapes, shown in *Figure 2*. These are commonly used building shapes in office design and represents architectural variations at an early stage of design. The first three shapes rectangular (*Shape1*), plus (*Shape2*) and L-shaped (*Shape3*) building are used for training CBML and the performance of the model is tested on H-shaped (*Shape4*) building. There are three options for the number of floors i.e. one floor, two floors and three floors and 2000 training instances for each of these options. Thus, there a is total of 18000 training instances, 6000 for each shape. *Shape4* mentioned under test data set is explicitly used for estimating and reporting the accuracy of developed ML models. The detail of design parameters and their ranges are mentioned in *Table 1*. We used Sobol sequence to generate a random combination of design parameters considering uniform distribution in the specified range (Herman & Usher, 2017).

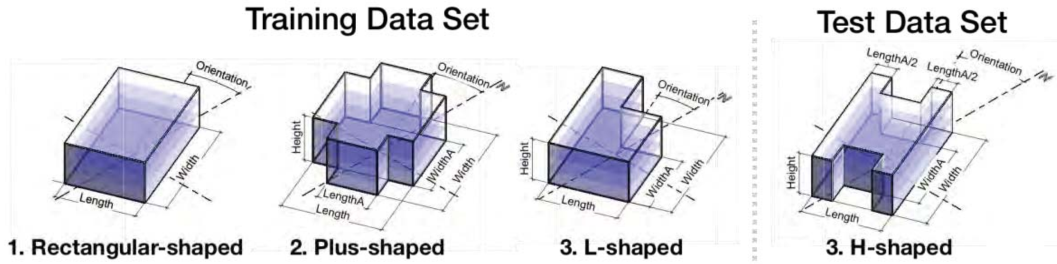

Figure 2. Architectural designs studied in this paper

Table 1. Detail of parameters with the ranges

Parameter	Unit	Range
Length	(meters)	10, 100
Width	(meters)	10, 100
Height	(meters)	3, 5
Orientation	Degrees	0, 180
Infiltration	(ACH)	0.2, 1
U_Wall	(W/m^2°K)	0.1, 0.75
U_GFloor	(W/m^2°K)	0.1, 0.75
U_Roof	(W/m^2°K)	0.1, 0.5
U_IFloor	(W/m^2°K)	0.1, 0.75
HC_Slab	(J/m^3°K)	800, 1600
U_Window	(W/m^2°K)	0.25, 1.5
g_Window	-	0.1, 0.9
WWR_N/E/W/S*	-	0.01, 0.95
Operating Hours	(hours)	8, 10
Lighting Heat Gain	W/m^2	5, 15
Equipment Heat Gain	W/m^2	10, 15
Chiller COP	-	3, 5
Boiler Efficiency	-	0.7, 0.9

A parametric simulation model has been set up with dynamic energy simulation tool EnergyPlus to generate data for training ML models. We have used weather data of Munich which represents most part of western Europe. The use of the building is office which follows a typical 5-days schedule. The parameters *U_Wall, U_GFloor, U_Roof, U_IFloor,* and *U_Window* imply u-value for walls, ground floor, roof, intermediate floors and windows respectively. *g_Window* implies g-value for windows and *HC_Slab* is heat capacity for floor slabs. *WWR_N/E/W/S* stands for window-to-wall ratio in north, east, west and south directions respectively and chiller COP is the coefficient of performance for chiller. The heating and cooling setpoints are 20 and 24°C and setback points are 10 and 28°C respectively. The energy model considers the effect of daylight in the zone and reduces the amount of artificial light to achieve lux level of 500. The occupant load is one person per 10 m^2. The simulation model is based one-zone-per-floor rule i.e. assuming one zone is present at each level.

The energy simulation is performed at the Vlaams Super Computer (VSC) using ten nodes equivalent to 360 cores at a clock speed of 2.3GHz.

4.2 Effect of enriching training data

Each ML model of CBML are trained using the first shape from training data i.e. rectangular building shape and the performance of ML models is used as a base case. To see the effect of increasing data, each model of CBML is trained again with additional shape one-by-one. Thus, recording the accuracy of each model of CBML with increasing data step-by-step will show the effect of enriching training data. To see the effect of training data, raw features listed in *Table 2* are used.

4.3 Effect of enhancing features

We are using two types of features: (1) *Raw Features* and (2) *Enhanced Features*. *Raw Features* are the features easily available for each component in digital models and mostly represents geometric and thermo-physical properties of the element. As listed in *Table 2*, *Raw Features* for building element level models are area and thermal heat transfer coefficients (u-value). *Raw Features* for zone level models represent size (floor area and height), heat flows, infiltration and internal heat gains. *Enhanced Features* for building element level model are *Raw Features* and additional features. For example, heat flow through wall component depends on the area, orientation and u-value. Additionally, radiation and the zone which is it associated must be represented in the features to predict heat flows. The zone is represented by *Zone Features* i.e. *Zone Area, Zone Volume, Total Light Heat Gain, Total Equipment Heat Gain, Total Infiltration, Operating Hours, Heat Capacity, Solar Radiation, Wall Area × U_Wall, GFloor Area × U_GFloor, Roof Area × U_Roof, Window Area × U_Window, IFloor Area × U_IFloor. Zone Features* and *Total Heat Flows* are used as input for *Zone Heating* and *Cooling Energy* model. Zone area, lighting or equipment heat gain and operating hours are used as a feature for lighting and equipment energy models. For *Building Total Energy*, both features represent the size of the building, system efficiency and energy outputs from zone level models.

Table 2. Details of features used in training each model of CBML

	ML Model	Raw Features	Enhanced Features
Building element level	**Wall**	area, orientation, u-value	area, orientation, u-value, radiation, zone features
	Window	area, orientation, u-value, g-value	area, orientation, u-value, g-value, radiation, zone features
	GFloor	area, u-value	area, u-value, zone features
	Roof	area, u-value	area, u-value, radiation, zone features
	IFloor	area, u-value	area, u-value, zone features, adjacent zone features
Zone level	**Heating energy**	area, height, light heat gain, equipment heat gain, infiltration, operating hours, heat capacity, total heat flows	zone features, total heat flows
	Cooling energy		
	Lighting energy	area, light heat gain, operating hours	area, total light heat gain, operating hours
	Equipment energy	area, equipment heat gain, operating hours	area, total equipment heat gain, operating hours
Building level	**Total energy**	total floor area, total volume, boiler efficiency, chiller COP, lighting energy, equipment energy, heating energy, cooling energy	total floor area, total volume, boiler efficiency, chiller COP, lighting energy, equipment energy, heating energy, cooling energy

We are using deep learning neural network architecture for training ML models which is a black-box estimator. To understand the importance of features in such ML model, we used permutation importance technique described in Breiman, 2001 and implemented using python library developed by Korobov & Lopuhin, 2016. The calculation of feature importance using the permutation importance technique is an iterative process. We have used 25 iterations to get reliable results. This step will give the feature importance of each feature in ML models and thus useful to know how enhancing the feature can be useful for further development.

5. Results

In this section, the prediction accuracy of CBML is documented to observe the effect of enhancing training data and features. The effect of enhancing training data is presented in section 4.1 and the effect of enhancing features is presented in section 4.2.

5.1 Effect of increasing training data

CBML is initially trained using the simulation data from *Shape1* (rectangular-shaped building). In the next step, the training data is enriched with the simulation data from *Shape2* (plus-shaped building) and finally, the simulation data from *Shape3* (L-shaped building) is also used for the training of ML models. The performance of each ML model of CBML, recorded in terms of R^2 and MAPE, is presented in *Figure 3*. *Total Heat Flows* is the sum of heat flows through the building components for a zone. The point (+) represents the performance of the ML model, the solid line represents the trends in the performance with respect to the number of shapes and the dash-dot line represents the linear-fit of the trend line. The performance of *Window* and *IFloor Heat Flow* models is not plotted in this figure as it will not fit in the scale. *Table 3* should be referred for the prediction accuracy of these component models. In most of ML models, the accuracy increases with the increase in training data, but it is not always the same. There is not much improvement in R^2 values for zone level or building level models as the values are quite high with *Shape1* also. But the MAPE improves for these models with the increase in the data, showing improvement in low-range values.

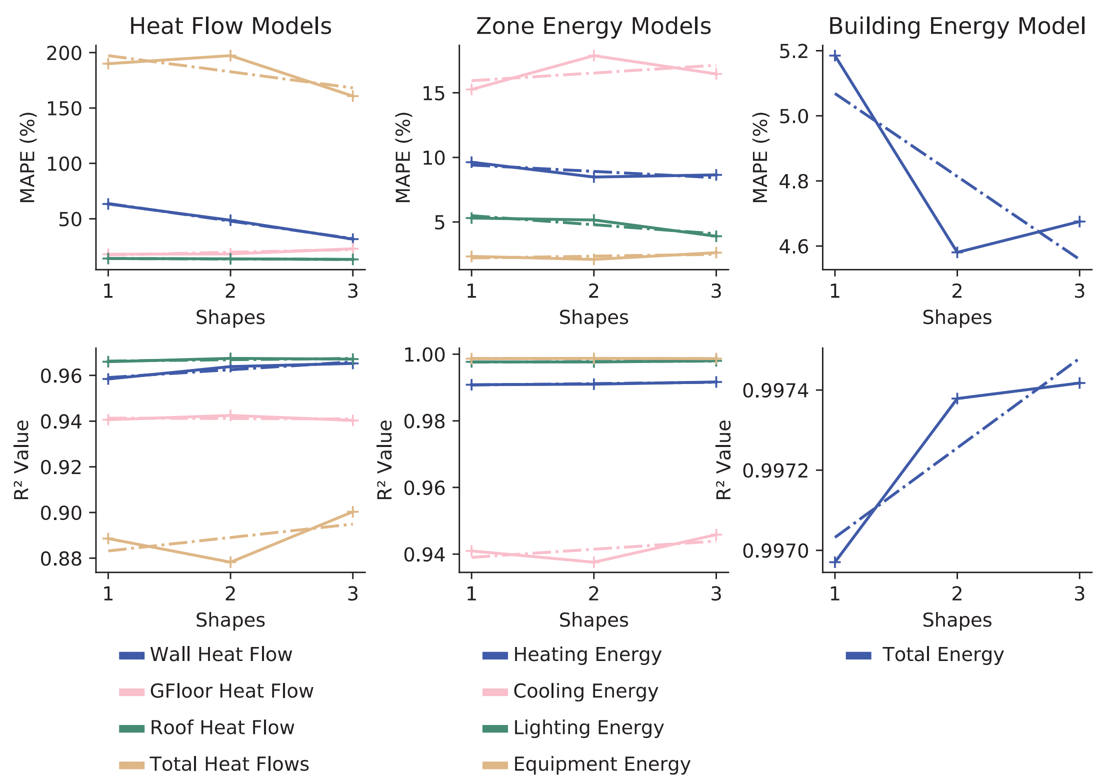

Figure 3. Effect of increasing training data on the prediction performance of ML models

5.2 Effect of enhancing features

Each model of CBML is trained using two types of features – raw and enhanced features as mentioned in *Table 2*. The performance of each ML model is recorded in terms of MAPE and R^2 and presented in

Table 3. It has three training scenarios of CBML. First, *Raw Features (Shape1)*, which means the training data is used from *Shape1* using *Raw Features*. Second, *Raw Features (Shape1-3)* means the training data is used from *Shape1, Shape2* and *Shape3* using *Raw Features* again. Third, *Enhanced Features (Shape1-3)* implies the same training data as the previous step but with *Enhanced Features*. It should be noted that there is a significant improvement in MAPE for each ML model. The accuracy of *IFloor* and *Window Heat Flow* model improves with the use of enhanced features. The accuracy of zone level models improves much more with the use of enhanced features compare the increasing in the training data. There is a similar improvement in the building energy model after increasing the training data or enhancing the features.

Table 3. Effect of training data and features on prediction accuracy of CBML

Feature and Shapes	**Raw Features (Shape1)**		**Raw Features (Shape1-3)**		**Enhanced Features (Shape1-3)**	
Model	**MAPE (%)**	**R2**	**MAPE (%)**	**R2**	**MAPE (%)**	**R2**
Wall Heat Flow	63.41	0.9584	31.68	0.9652	22.97	0.9889
Window Heat Flow	18.05	0.9406	814.03	0.9745	328.61	0.9922
GFloor Heat Flow	14.09	0.9660	22.96	0.9403	8.95	0.9882
Roof Heat Flow	18.67	0.9649	13.30	0.9671	7.70	0.9888
IFloor Heat Flow	340.46	0.0145	349.16	0.0143	71.24	0.9885
Total Heat Flows	189.97	0.8886	160.82	0.9003	91.73	0.9863
Zone Heating Energy	9.623	0.9908	8.64	0.9917	4.18	0.9981
Zone Cooling Energy	15.26	0.9410	16.46	0.9459	7.75	0.9838
Zone Lighting Energy	5.30	0.9977	3.90	0.9980	4.60	0.9978
Zone Equipment Energy	2.33	0.9987	2.62	0.9986	1.69	0.9989
Building Total Energy	5.18	0.9970	4.68	0.9974	3.14	0.9988

The feature importance is calculated for the last model only i.e. *Total Energy Model* trained using *Shape1-3* and *Enhanced Features*. The results for each model are presented as bar graphs in *Figure 4***Error! Reference source not found.**. The *Zone Features* consists of several features representing the characteristics of zone. Hence the feature importance of *Zone Features* is sum of the feature importance of these features. For *Wall* and *Window* heat flow, the area is the most important feature. For *Roof* and *GFloor*, *Zone Features* is the most important feature, which means there is some feature in *Zone Features* which characterizes the heat flow through these components more appropriately. For *IFloor, Zone Features* and *Adjacent Zone Features* shows similar importance. For *Zone Heating Energy* model, *Zone Features* are quite important compared to *Total Heat Flows* which represents heat flow through the building elements. For *Zone Cooling Energy* model, both *Total Heat Flows* and *Zone Features* are important. For *Zone Lighting* and *Equipment Energy* models, the results are self-explanatory. For *Building Total Energy* model, *Heating Energy* has the highest influence, followed by *Equipment Energy*, *Lighting Energy*, *Total Floor Area* and *Boiler Efficiency*.

Figure 4. Feature importance for each model (enhanced features are marked with red colour)

6. Discussion

The section discusses the applicability and limitations of trained ML models in the wider context. CBML offers more generalizable structure to perform energy prediction for the building shapes of architectural complexity. The developed ML models are performing with good accuracy on the design case outside of the training data set from the perspective of the whole building with its specific architectural shape. The advantage of components is that their usage stays within the training range. As a consequence, these models are useful for the energy prediction of building shapes which are not included in the training case.

It is expected that the prediction accuracy of the ML model will improve with as the training data increases, but not always. *Building Total Energy* model is performing better when the training data of *Shape1* is enriched with the data of *Shape2*, but there is no improvement when it is enriched further with the data of *Shape3*. Also, there is not much improvement in R^2 for total energy model which is close to 0.999 in the first instance itself. It means that the data enrichment is useful to improve the accuracy for low range values. The building element level models always show improvement after data enrichment. For zone level models, sometimes it performs better and sometimes there is no improvement. There is further investigation required to when enriching data with a new shape will be useful.

The feature enhancement improves the prediction accuracy of each ML model of CBML. It will be useful to know which feature influences the accuracy of the ML model as adding all the features may not be useful. Furthermore, it provides useful information for the selection of features. It is evident from *Figure 3* that the inaccuracies in ML models at building element level may or may not result in the inaccuracies in zone level models. It depends on the importance of heat flows in the zone level models. The feature importance for zone level models reveals that the heat flows only influence the *Zone Cooling Energy Model* a little and has no almost influence on the *Zone Heating Energy Model*. Also*, Total Energy* model at the building level, *Heating Energy* has the highest influence which is not influenced by *Total Heat Flows*. Thus, the inaccuracies in building element level model result in less inaccuracies in zone level models or building level model.

The training of CBML is performed using the data generated using dynamic energy simulation tool. The data generation process is simplified following certain assumptions such as *one-zone-per-floor model*, no presence of urban context and use of building as office etc. Thus, the CBML is valid for the assumptions which are used to perform dynamic energy simulation of buildings. Also, it utilizes weather data for Munich. However, a similar approach can be adopted after generating data with other assumptions and location.

7. Conclusions

The component-based machine learning model is useful for extending the ML model developed with some building shapes on the building shapes outside of the training data. The study shows that the prediction accuracy of the model improves with the inclusion of more building shapes in the training data in general. Also, it has been shown that CBML can be used to make the energy prediction of a building shape which is not included in the training data. The accuracy of total energy model at building level improves from 5.18% to 4.68 % *(MAPE)* and 0.9970 to 0.9974 *(R^2)* after enriching the training data with *Shape2* and *Shape3*. However, it is complex to understand which building shapes should be used for data enrichment.

The prediction accuracy of each model in CBML improves after enhancing the features. The feature importance exercise confirms that the additional features influence the prediction accuracy of ML models. For building element level models, *Zone Features* are really important as after improvement, the accuracy of models improve significantly evident with the reduction in MAPE and increase in R^2 values. The similar trend follows for zone level models as both performance measures improve after enhancing the features. This is evident in the feature importance graph also where the additional features show good importance in the zone level models. We didn't enhance the features for total energy model at the building level, but more accurate prediction at building element level and zone level models improves its prediction accuracy. The accuracy of total energy model at building level improves from 4.68% to 3.14 % *(MAPE)* and 0.9974 to 0.9988 *(R^2)* after enhancing the features. But the features should be selected more carefully for the development of each model of CBML as all the additional features may not be important to improve the prediction accuracy. Thus, the feature enhancement should be supplemented with the calculation of feature importance for each model. It is useful for the identification of relevant features for each model. The overall accuracy of *Total Energy* model at building level improves from 5.18% to 3.14% *(MAPE)* and 0.9970 to 0.9988 *(R^2)* after enriching training data and enhancing the features.

Acknowledgement

The authors want to acknowledge the support of Deutsche Forschungsgemeinschaft (DFG) for funding the research through research unit FOR 2363. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

References

Abualdenien, J., & Borrmann, A. (2018). Multi-LOD model for describing uncertainty and checking requirements in different design stages. In *Proceedings of the 12th European Conference on Product and Process Modeling*. Copenhagen, Denmark.

Asadi, S., Amiri, S. S., & Mottahedi, M. (2014). On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design. *Energy and Buildings*, *85*, 246–255. https://doi.org/10.1016/j.enbuild.2014.07.096

Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 26–33). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073017

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Catalina, T., Virgone, J., & Blanco, E. (2008). Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and Buildings*, *40*(10), 1825–1832. https://doi.org/10.1016/j.enbuild.2008.04.001

Cheng, M.-Y., & Cao, M.-T. (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing*, *22*, 178–188. https://doi.org/10.1016/j.asoc.2014.05.015

Chollet, F. (2015). Keras. Retrieved May 6, 2019, from https://keras.io

Fumo, N. (2014). A review on the basics of building energy estimation. *Renewable and Sustainable Energy Reviews*, *31*, 53–60. https://doi.org/10.1016/j.rser.2013.11.040

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. (N. Tache, Ed.) (Ninth). Sebastopol, CA: O'Reilly Media, Inc. Retrieved from https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/

Geyer, P., & Singaravel, S. (2018). Component-based machine learning for performance prediction in building design. *Applied Energy*, *228*, 1439–1453. https://doi.org/10.1016/j.apenergy.2018.07.011

Geyer, P., Singh, M. M., & Singaravel, S. (2018). Component-Based Machine Learning for Energy Performance Prediction by MultiLOD Models in the Early Phases of Building Design. In *Advanced Computing Strategies for Engineering* (pp. 516–534). https://doi.org/10.1007/978-3-319-91635-4_27

Halevy, A., Norvig, P., & Pereira, F. (2009). *The Unreasonable Effectiveness of Data*. *Intelligent Systems, IEEE* (Vol. 24). https://doi.org/10.1109/MIS.2009.36

Herman, J., & Usher, W. (2017). SALib: An open-source Python library for Sensitivity Analysis. *The Journal of Open Source Software*, *2*(9), 97. https://doi.org/10.21105/joss.00097

Korobov, M., & Lopuhin, K. (2016). Inspecting Black-Box Estimators. Retrieved April 10, 2019, from https://eli5.readthedocs.io/en/latest/blackbox/index.html

Ng, A. Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Twenty-first international conference on Machine learning - ICML '04* (p. 78). New York, New York, USA: ACM Press. https://doi.org/10.1145/1015330.1015435

Singaravel, S., Geyer, P., & Suykens, J. (2018). Deep Learning Neural Networks Architectures and Methods: Building Design Energy Prediction by Component-Based Models. *Advanced Engineering Informatics*, submitted.

Singh, M. M., Singaravel, S., & Geyer, P. (2018). Information Exchange Scenarios between Machine Learning Energy Prediction Model and BIM at Early Stage of Design. In R. Caspeele, L. Taerwe, & D. M. Frangopol (Eds.), *Publications of the Sixth International Symposium on Life-Cycle Engineering* (pp. 487–494). Ghent, Belgium: CRC Press.

Struck, C., de Wilde, P. J. C. J., Hopfe, C. J., & Hensen, J. L. M. (2009). An investigation of the option space in conceptual building design for advanced building simulation. *Advanced Engineering Informatics*, *23*(4), 386–395. https://doi.org/10.1016/j.aei.2009.06.004

Tian, W., Heo, Y., de Wilde, P., Li, Z., Yan, D., Park, C. S., … Augenbroe, G. (2018). A review of uncertainty analysis in building energy assessment. *Renewable and Sustainable Energy Reviews*, *93*, 285–301. https://doi.org/10.1016/j.rser.2018.05.029

Van Gelder, L., Das, P., Janssen, H., & Roels, S. (2014). Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. *Simulation Modelling Practice and Theory*, *49*, 245–257. https://doi.org/10.1016/j.simpat.2014.10.004

Van Gelder, L., Janssen, H., & Roels, S. (2014). Probabilistic design and analysis of building performances: Methodology and application example. *Energy and Buildings*, *79*, 202–211. https://doi.org/10.1016/j.enbuild.2014.04.042