Using surrogate models to analyze the impact of geometry on the energy efficiency of buildings

by

Bhumika Bhatta
B.Arch, Tribhuvan University, 2018

A thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Applied Science

in the Department of Civil Engineering

Using surrogate models to analyze the impact of geometry on the energy efficiency of buildings

by

Bhumika Bhatta
B.Arch, Tribhuvan University, 2018

Supervisory Committee

Dr. Ralph Evins, Supervisor
(Department of Civil Engineering)

Dr. Christopher Kennedy, Departmental Member
(Department of Civil Engineering)

ABSTRACT

In recent times data-driven approaches to parametrically optimize and explore building geometry has been proven to be a powerful tool that can replace computationally expensive and time-consuming simulations for energy prediction in the early design process. In this research, we explore the use of surrogate models, i.e. efficient statistical approximations of expensive physics-based building simulation models, to lower the computational burden of large-scale building geometry analysis. We try different approaches and techniques to train a machine learning model using multiple datasets to analyze the impact of geometry and envelope features on the energy efficiency of buildings. These contributions are presented in the form of two conference papers and one journal paper (being prepared for submission) that iteratively build up the underlying methodology.

The first conference paper contains preliminary experiments using 4 manually generated building geometries for office buildings. Data were generated by simulating various building samples in EnergyPlus for different geometries. We used the generated data to train a machine learning model using support vector regression. We trained two separate models for predicting heating and cooling loads. The lesson learned from this first experiment was that the prediction of the models was not great due to insufficient geometric features explaining the variability in geometry and the lack of sufficient data for varied geometries.

The second conference paper developed a novel dataset of 38,000 building energy models for varied geometry using 2D images of real-world residences. We developed a workflow in the Grasshopper/Rhino environment which can convert 2D images of a floor plan into a vector format then into a building energy model ready to be simulated in EnergyPlus. The workflow can also extract up to 20 geometric features from the model, to be used as features in the machine learning process. We used these features and the simulation results to train a neural network-based surrogate model. A sensitivity analysis was performed to understand the impact and importance of each feature to the energy use of the building. From the results of the experiment, we found that off-the-shelf neural network-based surrogates provided with engineered features can very well emulate the desired simulation outputs. We also repeated the experiment for 6 different climatic zones across Canada to understand the impact of geometric features across various climates; these findings are presented in an appendix.

In the journal paper, we explored two different methodologies to train surrogate models: monolithic and component-based. We explored the component-based modeling technique as it allows the model to be more versatile if we need to add more components to it, ultimately increasing the usability of the model. We conducted further experiments by adding complexity to the geometry surrogate model. We introduced 10 envelope features as an input to the surrogate along with the 20 geometric features. We trained 6 different surrogate models using different datasets by varying geometric and envelope features. From the results of the experiment, we found that the monolithic model performs the best but the component-based surrogate also falls into an acceptable range of accuracy.

From the overall results across the three papers, we see that simple neural network-based surrogate models perform really well to emulate simulation outcomes over a wide variety of geometries and envelope features.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Ralph Evins, for helping and supporting me throughout my degree and encouraging me in every step where I had doubts. Additionally, I would also like to thank my fellows at Energy in cities for making my master's journey enjoyable.

I would also like to thank my family for their unconditional love and support. Special thanks to my partner Sushil for inspiring me, listening to all my rants, and helping me remain sane throughout the journey.

Author Contribution

This thesis consists of two peer-reviewed conference papers and one journal manuscript that is ready for submission to the Journal of Building Performance Simulation. Full citations and author contribution details for each work are given below.

1. Paper: 1: Bhatta, B., Faure, G., Evins, R. (2021, June). *Using a surrogate model to analyze the impact of geometry on the energy efficiency of buildings with optimal envelope design.* eSim 2021, Vancouver, Canada chapter(2).
   B.B. developed the methodology, performed the analysis, and wrote the manuscript. G.F contributed to the methodology and analysis. R.E. supervised the project and revised the manuscript.

2. Paper: 2: Bhatta, B., Westermann, P., Evins, R. (2021, August). *Using a surrogate model to analyze the impact of geometry on the energy efficiency of buildings.* Building Simulation 2021, Bruges, Belgium (Chapter3).
   B.B. developed the methodology, performed the analysis, and wrote the manuscript. P.W contributed to the methodology. R.E. supervised the project and revised the manuscript.

3. Paper: 3: Bhatta, B., Westermann, P., Evins, R. (2021). *Comparison of monolithic and component-based surrogate models to analyze the impact of geometry for energy efficiency in the building.* To be submitted to the Journal of Building Performance Simulation chapter (4).
   B.B. developed the methodology, performed the analysis, and wrote the manuscript. P.W contributed to the methodology. R.E. supervised the project and revised the manuscript.

Chapter 1

Introduction

Buildings cover a major share of global energy consumption and have a large impact on the environment [37]. There are multiple factors that affect the energy efficiency in a building such as construction, form, design, systems, controls, etc [66]. Among these properties, building geometry in particular has a great effect on energy performance [30]. Research works such as [21], [50], [44] show that building shape has impact on heating and cooling load in a building. Unlike other building properties, building geometry is a feature that usually remains constant throughout the lifecycle of the building [30],[27],[23]. This fact makes it important for an architect or a designer to analyze the impact of the shape or design of the building they choose in the conceptual phase. However, there are no general guidelines or rules of thumb available to an architect to analyze the impact of the shape they choose.

There are various tools developed for building information modelling (BIM) and building energy modelling (BEM) [14] such as EnergyPlus [19], TrnSys [29], ESP-r [61], and Modelica [68]. However, BIM and BEM can be time-intensive steps. Hence, detailed exploration of geometry at the early design phase remains an unconventional, expensive, and time-consuming procedure. The building industry lacks a tool that can provide a quick and easy assessment of energy consumption for geometry exploration. In order to bridge this gap, we have developed machine learning models, also known as surrogate models, for building geometry and envelope features to guide early-stage decision-making.

Surrogate models are statistical models developed by using machine learning techniques that can provide a rapid approximation for a detailed simulation in a relatively small time which helps in reducing the computational barriers in building energy simulation [66]. Figure 1.1 shows the process of surrogate modelling. The steps for

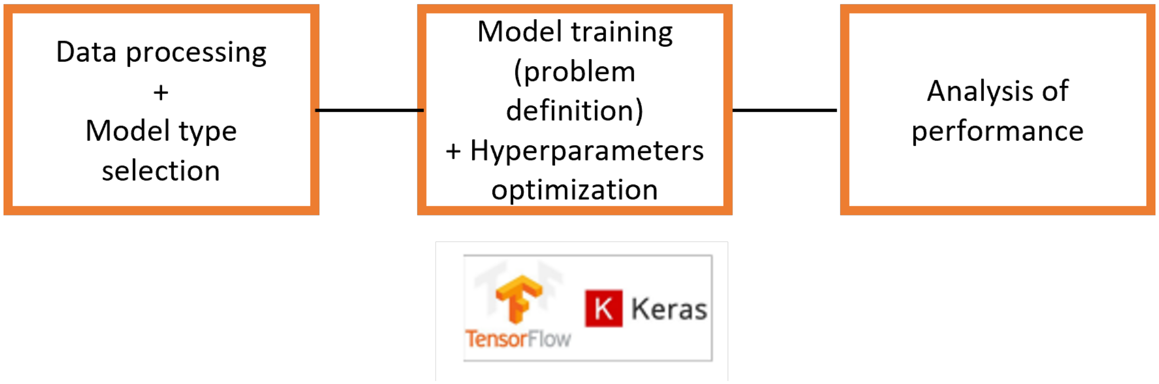


Figure 1.1: Surrogate model training procedure

surrogate model training are :

- Data processing and model type selection. We need to select or generate appropriate and usable data format to train the model and select a particular type of meta-modelling techniques for regressions such as SVR, Neural networks, etc.
- Problem definition (Input and objective), model training, and hyperparameter optimizations.
- Validating the model's accuracy using different metrics such as R^2, MSE, RMSE, etc.

This approach can provide a rapid approximation of energy use for a building without the burden of simulating each and every design idea in the conceptual phase. Surrogate models have shown promising results to provide a building performance assessment faster and less computationally expensive than building simulation-based performance analysis [66].

A lot of advances have been made recently in data-driven approaches to substitute traditional computationally expensive and time extensive building performance simulations. The data-driven approaches allow the development of quick energy prediction tools which helps architects and designers integrate energy performance analysis in the early design phase [59]. At the early stage, approximations of building energy performance are used to compare and analyze various designs options, which is more useful than a detailed energy consumption prediction of a single design option. The approach we developed in this research work facilitates quick energy predictions making it easy to compare various designs in the conceptual phase without the burden of time-consuming simulations, though there is a small sacrifice in the accuracy in prediction.

In this research work, we have developed multiple surrogate models. Initially, we started with manually generated building geometries which were developed in OpenStudio. The shapes we chose for the preliminary experiments were rectangle, square, L-shaped, U-shaped, and T-shaped buildings. We used support vector regression (SVR) to train the surrogate model. From the preliminary results, we found that these surrogate models didn't have a good level of accuracy in predicting the energy use which might be because of less variation in the geometry, small dataset, and insufficient geometric features. For the next step, we wanted to increase the sample size of different geometries for the analysis. For this, we developed a workflow to convert 2D floor plans into building energy models in the Rhino/Grasshopper environment. We developed 38,000 building energy models of different geometries for residences by using a real-world data set of 2D images of the apartment buildings in China [69]. We trained a neural-network-based surrogate model on a data set that was generated by simulating 38,000 different building geometries. The performance of the model increased due to variation in geometry and better input for geometric features. For the third stage, we increased the complexity of the model by adding envelope features as input to the surrogate model along with geometric features. In the third paper 4, we also explored two different methodologies, monolithic and component-based surrogate modelling. Monolithic model training has the advantage of high accuracy and performs best for a particular type of analysis, but component-based surrogate models provide more flexibility as different design aspects of a building could be added as individual components in the surrogate model, making the whole building design analysis much easier and convenient. We compared the monolithic and component-based surrogate modelling technique by training 6 different surrogate models with varied datasets and compared their accuracy using three error metrics (R2 score, mean squared error, and root mean squared error).

1.1 Summary of contribution

1. Applying surrogate modelling to building geometry (Papers 1,2,3).
   We are the first to apply the surrogate modelling process to input variable that handle building geometry in a detailed manner. Prior works were limited to few inputs for geometry such as relative compactness, volume etc [50], [52], [49] etc.

2. Generating a novel dataset of 38,000 building energy models over a wide variety of geometry (Paper 2).

   we developed our own data set of 38,000 building geometry for the detailed geometry analysis. Most of the research work done in geometry are limited to basic shapes [39] etc.

3. Detecting important geometry and envelope parameters for building energy efficiency using sensitivity analysis (Paper 1,2).

   In this work not only train a surrogate model for predicting building energy use but also do a sensitivity analysis to understand the importance of each selected features which provides a great insight for a designer at an early conceptual phase.

4. Applying monolithic and component-based surrogate modelling techniques for training the surrogate models (Paper 3).

   We use two different methods to train surrogate models. Prior research work has been limited to train models using monolithic surrogate modelling techniques [55],[36] etc.we are first to apply component based surrogate modelling techniques and compare it with the performance of monolithic model.

5. Analyzing the impact of geometry on various climatic zones (Extension of paper 2, (Appendix)).

   We analyze the impact of geometry on different climates by generating simulation data of various geometry for 6 different climate zones across Canada which makes the research more applicable.

Chapter 2

Using a Surrogate Model to Analyze the Impact of Geometry on the Energy Efficiency of Buildings (Paper:1)

Bhumika Bhatta[12], Gaelle Faure [12], Ralph Evins[12]
[1]Energy in Cities group, Department of Civil Engineering, University of Victoria, B.C, Canada.
[2]Institute for Integrated Energy Systems, University of Victoria, B.C, Canada.}

2.1 Abstract

Parametric optimization of building geometry has been demonstrated as a powerful technique to design energy efficient buildings. However, it remains in practice a computationally expensive and time-consuming process. Surrogates are the statistical models which can be used to provide rapid approximations of more expensive models. Surrogate models can be taken as an alternative approach to replace the detailed simulations with simplified approximate simulations, thereby sacrificing accuracy for reduced computational time.

In this research a surrogate model is used to approximate the impact of building geometry on energy demands. Data is generated from simulation results obtained from Energy Plus for different building shapes with varied envelope features. The envelope design is represented by a set of parameters such as window to wall ratio, aspect ratio, orientation, south wall area, total wall area, shape etc. The methodology is applied to a series of office buildings with different geometric shapes. Sensitivity analysis is performed to analyze the relationship between building features and energy use of the building. Data generated from the simulation process is used to train a machine learning model which approximates the heating and cooling loads for different shapes, thus acting as a surrogate model for the detailed simulation. This allows shape to be accounted for the optimization of building features without the burden of simulating and optimizing building design individually. The accuracy of the two surrogate models (for cooling load and heating load) were measured with two error metrics (R^2-score and MAPE score). The R^2-score for heating load predictions for the trained surrogate model is 0.965 and for the cooling load is 0.985 and the MAPE score of the models trained for cooling predictions is 0.49 and for heating predictions is 0.63.

2.2 Keywords

Surrogate Model, District Heating load, District cooling load, Building geometry, Machine learning.

2.3 Introduction

A Building is a function of its shape property, location etc, these parameters vary according to building type. Hence, the ways of providing energy efficiency also change. Therefore, a conscious approach needs to be developed in order to reach the optimum solution for energy efficiency in preliminary architectural design by exploring necessary data [72]. Energy performance of buildings relies hugely on form and geometry. The footprint of the building is decided at the initial phase of design but has a huge impact on the overall energy consumption of the building. However, very few tools are available for architects to analyze the effect of geometry on energy consumption. Many studies have been conducted to understand the relationship between building morphology and energy performance, but the relationship varies widely with location, climate and operational characteristics. Hence a building, when optimized in accordance to location, climate and function, could result in significant reduction of energy and operational cost [39]. The work presented in the paper analyzes the impact of

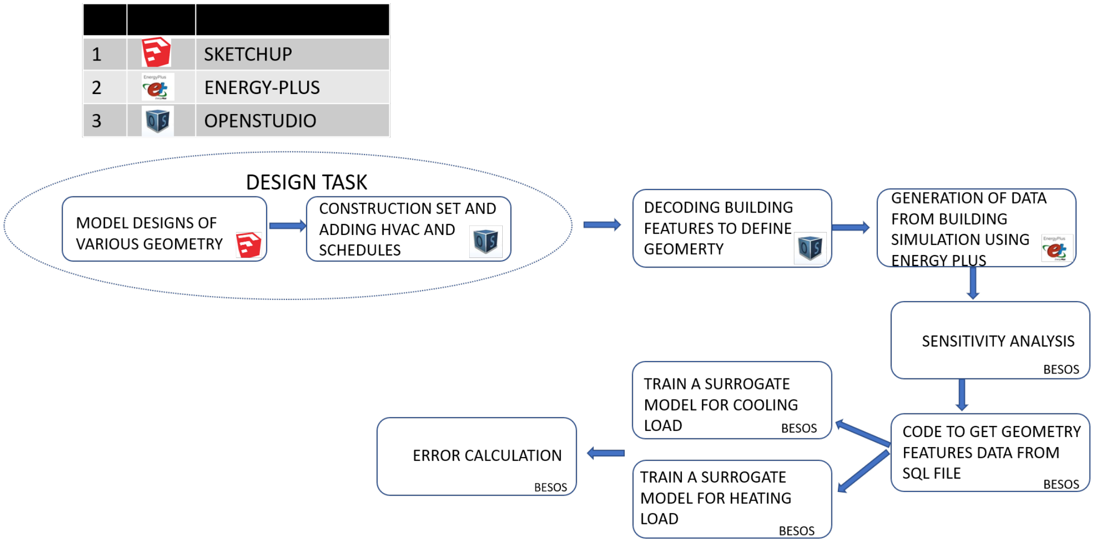


Figure 2.1: Flow-chart showing the workflow of research

building geometry on the total energy consumption of an office building by developing a surrogate model. Facilitating and automating the building design process will be crucial to increase the share of energy efficient buildings. Recent advances in machine learning are pushing the automation of problems related to energy efficient building design. [46],[42]. However, current building simulation software are computationally expensive and time intensive. Surrogate models are promising to provide building performance assessment much faster than simulation-based design analysis [66].

Many research studies have considered hypothetical building shapes with constant floor area and height or constant floor area with varying height among different shapes. Khamma et al. has considered the Department of Energy's (DOE) commercial reference building for their study [39]. The work presented in this paper uses a model, based on the "Building Technology Assessment Platform" (BTAP) small and medium office building for Victoria, BC. Other research work has been done on the building geometry and its impact on building energy. AlAnzi et al. takes the geometry of building into account for the energy performance analysis, but considers the shape of facade only and not the overall footprint [5]. Most research studies sought to identify the relationship between the relative compactness or building volume and building energy loads but does not look at other geometric features that might affect the energy efficiency of the building [49], [50], [52].
Building geometry can be illustrated by various simple numeric indicators such as as-

pect ratio, floor area, relative compactness, height of the building, volume, south wall area, total wall area, shape etc. Adams et al. uses 55 different features in-order to represent building morphology, fenestration, construction properties, internal gains, operating times, wind modifiers, flow paths, window control, and soil traits, in which 17 of the parameters are just for the building form layout and window geometry [51]. Adam's research is focused on developing a meta model to understand the performance of the design and simple evaluation of naturally ventilated commercial buildings in hot and humid climates. The work presented in our research has 9 different parameters which are used to define the building form and window geometry. The numeric indicators used to illustrate the building geometry are explained in Table 3.1. The models created were simulated in Energy plus [20]. The weather file used for simulation was for Victoria, BC [4]. The analysis was performed in BESOS, a cloud-based platform for building simulation and surrogate modelling [26]. The results the effect of numeric indicators on total heating and cooling loads in the building. Sensitivity analysis of the building features was performed in order to see the impact of each parameter. In the second part of the work, a surrogate model was trained with inputs of building features and output of heating and cooling loads. Two separate predictive models for heating and cooling loads were created and the error of the model were assessed using two error metrics [1], [3].

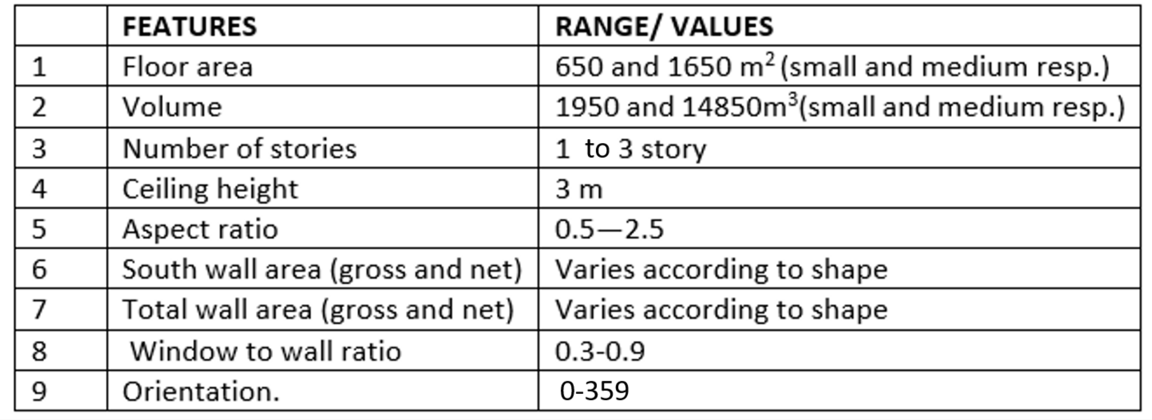
Building energy simulation is a computationally expensive procedure but building designers and architects requires fast results. The objective of this paper is to introduce a surrogate model which takes the building features as input and predicts the heating and cooling loads. This could eliminate the tedious process of building geometry definition and hence simplifies the work for the designer. The content of the paper is as follows: The first section gives the methodology and details of the models are given. The next part gives some discussion and conclusion.

2.4 Methodology

The work flow of the research is shown in Figure 2.1. First, the impact of geometric shape on the total heating and cooling loads of the building is analyzed by comparing 4 different shapes(Rectangle, U, T and L) and with 3 different aspect-ratios (1:1, 1:1.5, 1:2) resulting in 12 different building models which were again varied in terms of window to wall ratio, orientation and aspect ratio computationally to generate 600 models. The software used was Sketch-up and Open studio [31], [16].

The selection of features to define a geometry is a crucial aspect as the main goal was to select the features with which the basic geometry of any building could be defined. The list of numeric indicators used in this research is presented in Table 2.1

Table 2.1: Description of building features



	FEATURES	RANGE/ VALUES
1	Floor area	650 and 1650 m^2 (small and medium resp.)
2	Volume	1950 and 14850m^3(small and medium resp.)
3	Number of stories	1 to 3 story
4	Ceiling height	3 m
5	Aspect ratio	0.5—2.5
6	South wall area (gross and net)	Varies according to shape
7	Total wall area (gross and net)	Varies according to shape
8	Window to wall ratio	0.3-0.9
9	Orientation.	0-359

2.4.1 Constant building properties.

Identical construction specifications for climate zone 5B, template 189.1-2009 were used. The floor and roof in the Building surface are set to be adiabatic to make it independent of building height. For the internal and external loads: NECB defaults for occupancy (25 m^2/occ), lighting (8.8 W/m^2), electrical equipment (7.5 W/m^2) was set. The building was equipped with Variable Air Volume, fan-coil HVAC system with a boiler and chiller. Individual thermostats are set in each thermal zone.

2.4.2 Parameter variation

Two sets of models were prepared for small office type building and medium office type building. The aspect ratio of models ranged from (0.5-2.5). The aspect ratio quantifies the building's footprint in a ratio of length and width (x:y), which is given by equation 1:

$$Aspect - ratio = \frac{length(x)}{width(y)} \tag{2.1}$$

Generally, office buildings have a large glazing area for views day-lighting but such extensive use of glazing can result in increased energy consumption for heating

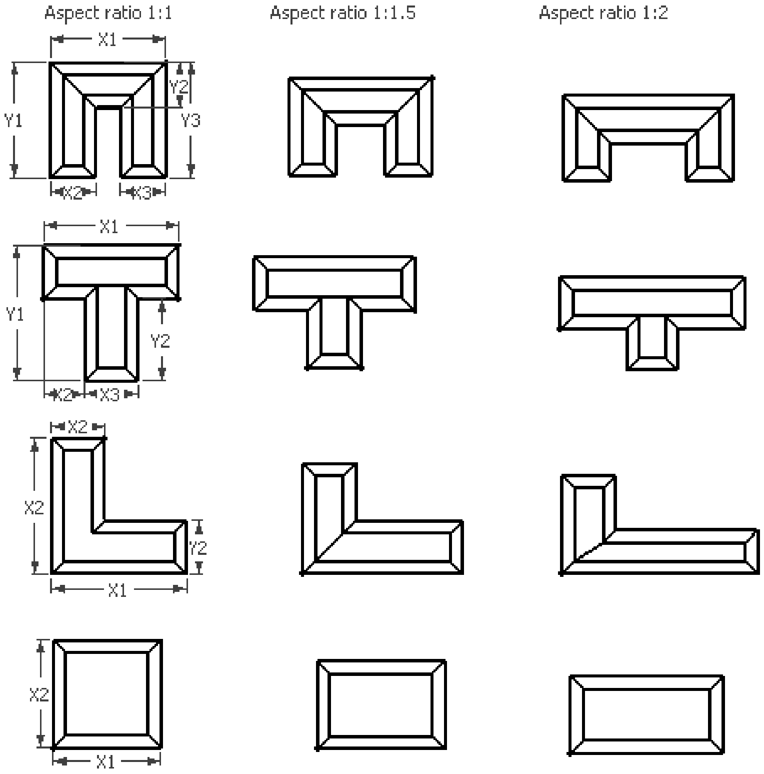


Figure 2.2: Building footprints of all the building models used in the research work.

and cooling. Hence, the orientation becomes one of the important deterministic factors. In this work the variation of the building orientation is varied between angles (0-359 degree).
Sensitivity analysis for all the building features which were not constant in each geometry (H,L,U,Rec) were plotted against heating and cooling loads to analyze the relationship. The features which were analyzed were aspect ratio, total wall area, south wall area, orientation, window to wall ratio (WWR) with the heating and cooling loads.

2.4.3 Surrogate modelling

Parametric optimization of building geometry has been demonstrated as a powerful technique to design energy efficient buildings, but it's not common in practice by the designers as it is a computationally expensive and time-consuming process. Also, designers lack the tools and knowledge for the optimization of various building parameters. The main aim of this paper is to use a surrogate model to approximate the energy demand of a building, which makes parametric optimization much more practical.

Support vector regression (SVR) [60] was used to train a surrogate model. SVR performs better than neural networks for small data sets [6]. The python library used for the SVR model is sci-kit learn. Since the research work is to predict two output (heating and cooling loads), two different SVR models were trained for heating and cooling loads. The surrogate models for heating cooling loads prediction had data for both small and medium office type buildings. The features of the building are total floor area, volume, ceiling height, number of stories, total wall area net and gross, south wall area gross and total, orientation, aspect ratio, shape and window to wall ratio. All the building features used as an inputs to the surrogate model were numeric except the shape feature, which was a categorical variable R, L, T, U (the starting letter of the shape) hence the shape feature was transformed into a column for each value i.e. one hot encoding [2]. Since the input values had different scales of values, they needed to be normalized for the better performance of the machine learning algorithm. The scaled values were used to train the model. The data was segregated into train and test data. 80% of the data was used to train the model whereas 20% of it was for testing the model.

2.5 Results

2.5.1 Sensitivity Analysis

The results are only analyzed for the small office building type as other building types have similar inherent patterns. From the plots in Fig 2.3 we can see that all the building points with higher value of energy loads are also high in WWR. Fig 2.3 shows the similar relationship pattern in all buildings for the south wall area, total wall area and window to wall ratio. With the increase in window to wall ratio the heating and cooling loads increases in each building. Similarly, decrease in heating load with increase in south wall area can be seen in all buildings, this might be because there is maximum solar heat gain. We can see that with increase in total wall area, the heating load of the building decreases.

The relationship between the cooling load and south wall area is not quite clear in Rec and U, but in T and L we can see a pattern that with increase in the south wall area there is decrease in cooling load. This might be because the south surface area at is less in L and T. Orientation of the building and the energy loads of the building seems to have a parabolic relation. Orientation with cooling load has a parabola facing upward whereas with heating load the parabola is facing downwards. This relationship must be considered while deciding the building orientation.

The total wall area of a building appears to be inversely proportional with the heating and cooling loads. To understand this relationship in more depth, the wall area for each shape with fixed floor areas were plotted Fig 2.7. The variation in the heating load for the same shape is due to other parameters such as orientation, WWR etc. Fig 2.7 shows the heating and cooling loads for 4 different shapes plotted with their wall area. We see that Rectangle has the least heating load as its surface area is the lowest but even if the surface area of U is the highest it does not have the maximum heating load instead, L has the highest heating load followed by T. This concludes that surface area alone is not a dominant factor for the heating load, other factors also come into play, for example self-shading angle of the building. The plot shown in Fig 2.6 doesn't show a pattern in the relationship of aspect ratio with the heating and cooling loads. The results are scattered and no particular shape is dominant. In such cases the surrogate model plays an important role as we can predict the energy loads of different aspect ratios in a very short period without running simulations for a variety of aspect ratios. The difference in heating and cooling loads of all the four shapes for small office buildings was also analyzed; the results can be seen in Fig 2.4. The plot shows four different clusters for the 4 different shapes. The rectangle shaped building has the lowest heating load but the maximum cooling load, U has a mid-range value among the four shapes for both heating and cooling loads, whereas L has the maximum heating load followed by T shaped building but both have the minimum cooling load. This plot can be helpful while deciding the building footprint.

2.5.2 Performance of Surrogate Modellings

The performance of the model for heating load(top) and cooling load(bottom) can be seen in fig 2.5. The two clusters are for the small office type and the medium office. The surrogate model for cooling load has more accuracy than that of the model for heating load.

2.5.3 Validation of results

We analysed the performance of the surrogate model using two different error metrics: R^2 score and MAPE.

- R^2 score is a statistical way to measure how close the data are to the fitted regression line. R^2 scores range between 0 and 100% where 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variance. The R^2-score for heating load predictions for the trained surrogate model is 0.965 and for the cooling load is 0.985. These are very respectable values showing that the model captures all but 3.5% and 1.5% of the variation respectively.

- The mean absolute percentage error (MAPE) is another statistical measurement of to identify the accuracy of a model.

  $$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{2.2}$$

  Where (A_t) is the actual values and (F_t) is the forecast values. It is given by equation (2):
  The MAPE of the models trained for cooling predictions is 0.49 and for heating predictions is 0.63.

2.6 Conclusion

The research in this paper presents a method that can help in the parametric simulation of geometry. Various numeric indicators were varied in small and medium office building models. The results from the energy simulation were used for a sensitivity analysis to see how the parameter affects heating and cooling loads. The same data was then used to train a surrogate model which used support vector for regression. Two different models were trained, each for the heating and cooling loads. Two error calculation metrics were used, which are R^2-score and MAPE. The R^2-score for heating load predictions for the trained surrogate model is 0.965 and for the cooling load is 0.985, which shows that the model captures all but 3.5% and 1.5% of the variation respectively and the MAPE score of the models trained for cooling predictions is 0.49 and for heating predictions is 0.63.

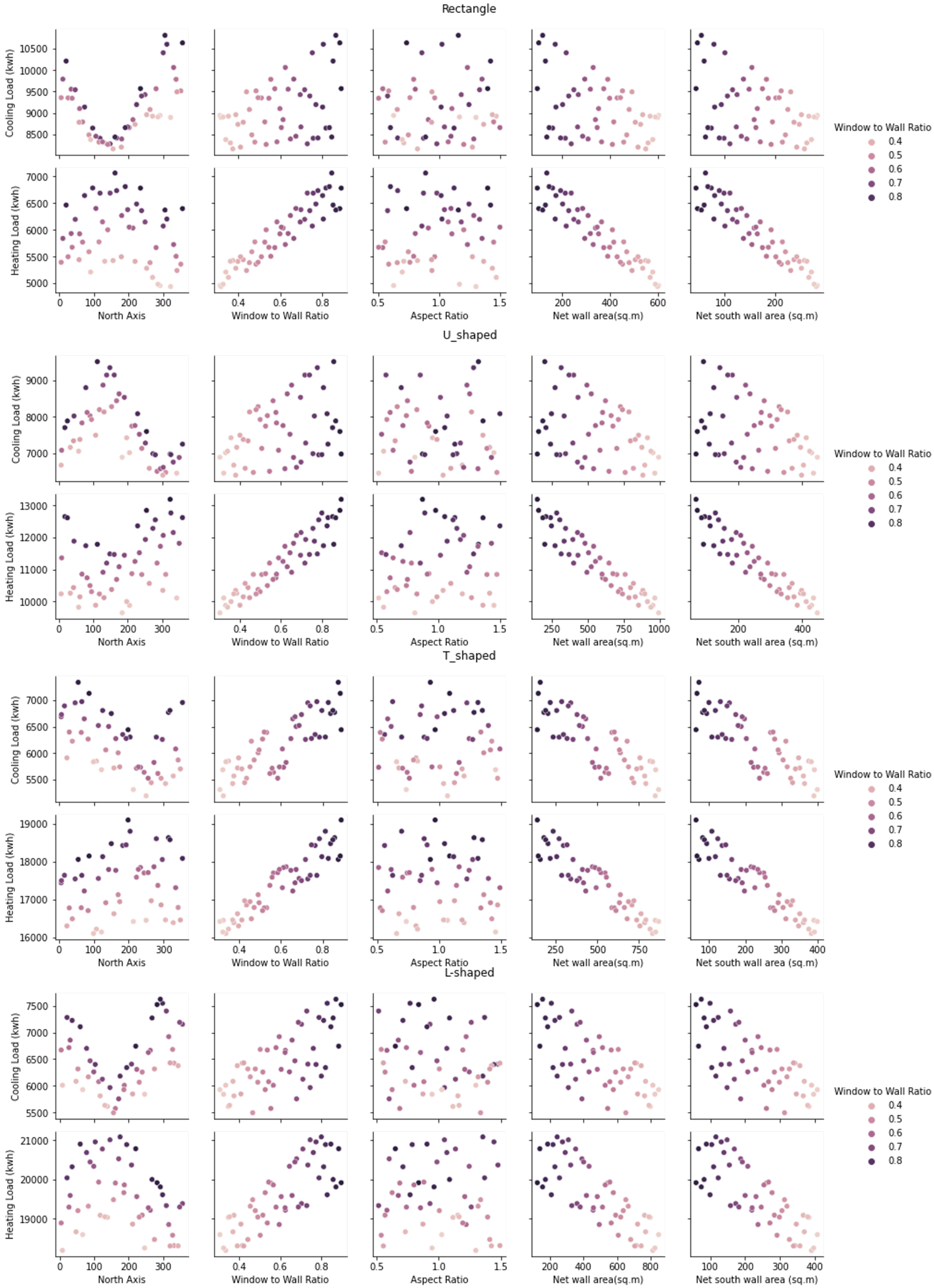


Figure 2.3: Sensitivity analysis of different shape features.

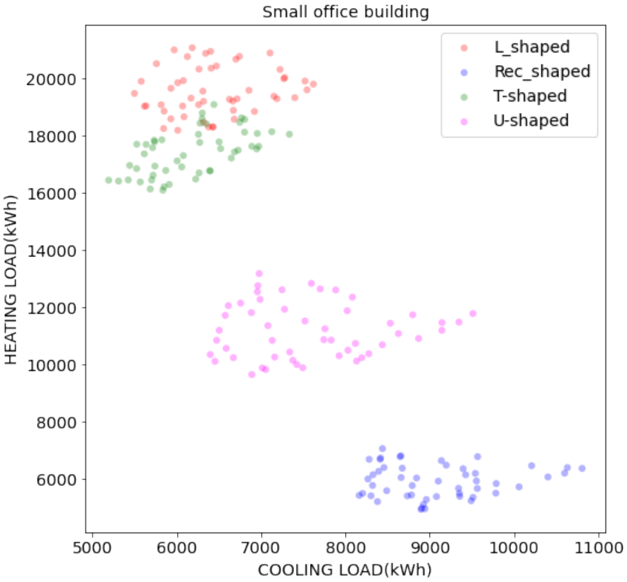


Figure 2.4: Plot of heating and cooling loads for all building shapes

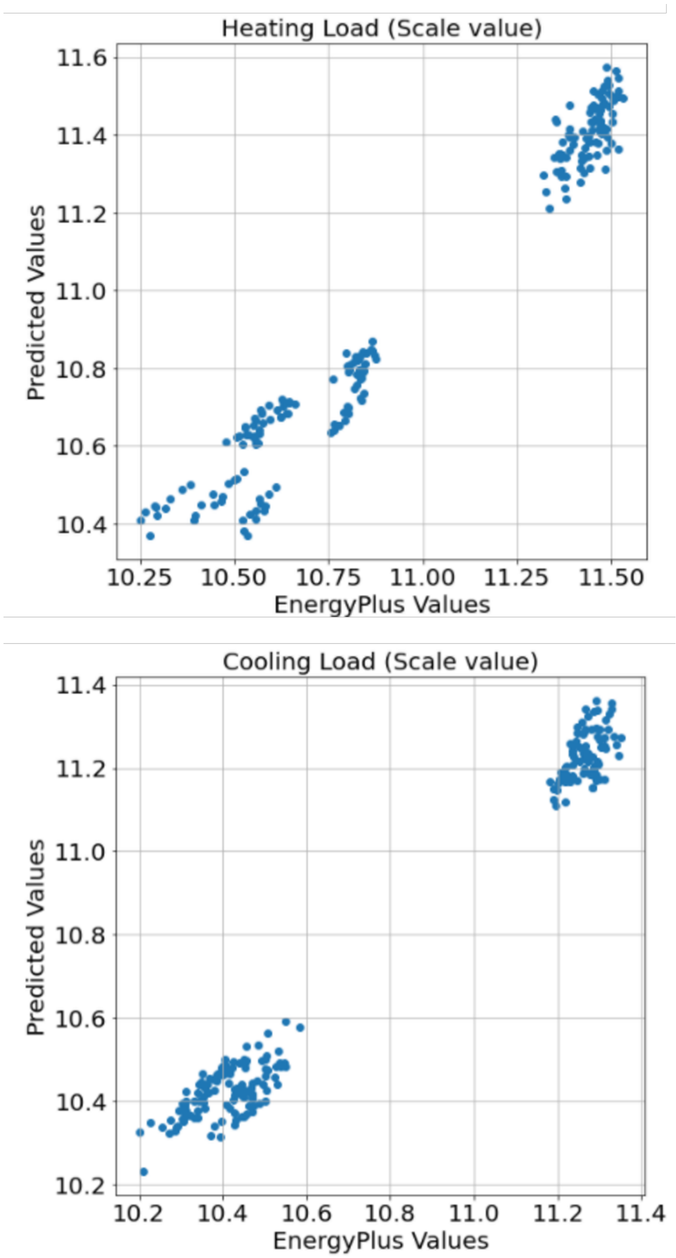


Figure 2.5: Plot of heating load prediction for the surrogate model

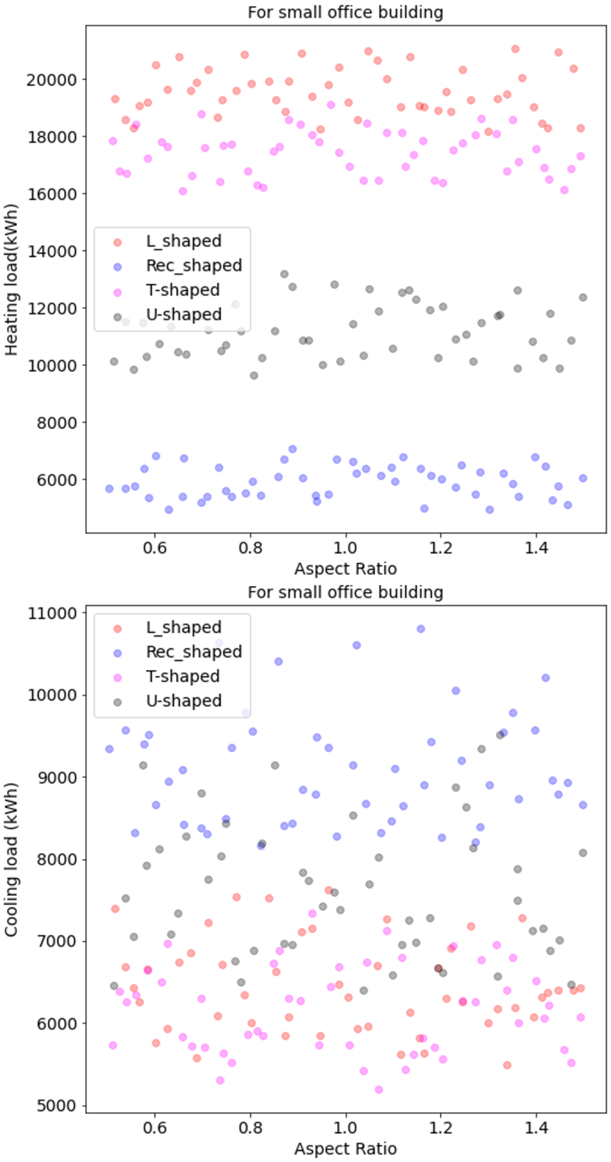


Figure 2.6: Relationship of aspect ratio with heating and cooling loads

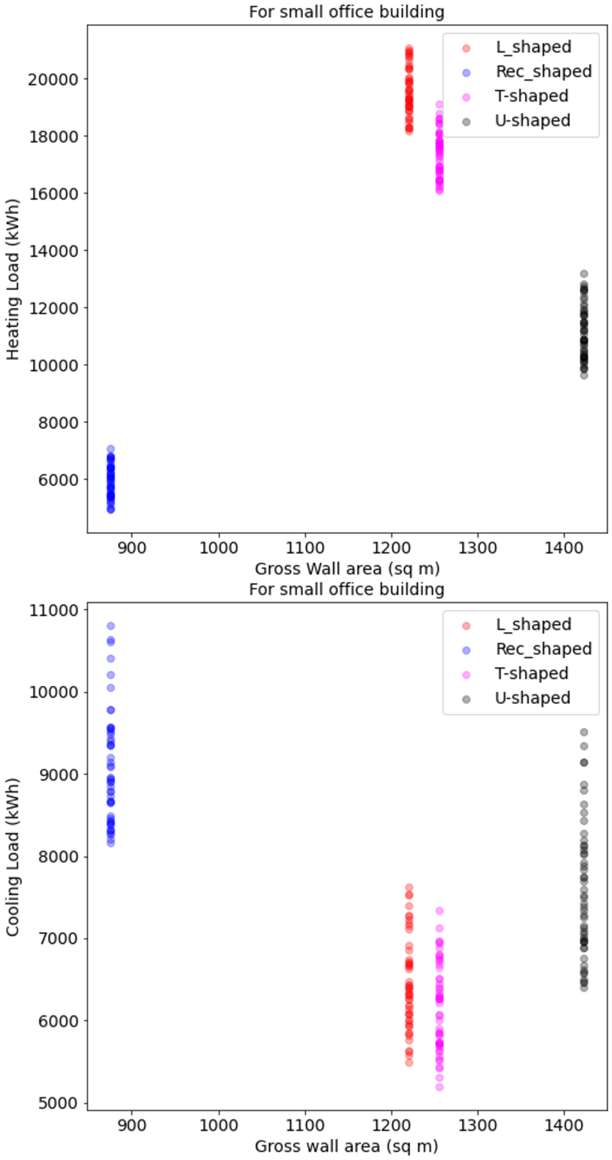


Figure 2.7: Plot of heating and cooling loads against total wall area

2.7 Discussion and future work

The results obtained from the graphs helps us understand the relationship of selected geometric features with the energy outputs. These results can be used as reference for early stage conceptual designs by guiding to choose optimum values for geometric parameters to achieve desired value of heating and cooling loads of the building. The current setup forms the basis for further research where the complexities will be increased by adding more geometric shapes and increasing the geometric features to define geometry.

From the results we see that some of the input features do not inherit a clear relationship with the output (heating and cooling loads), this suggests that there are other building geometry features that might be more important. Hence, the future work will be addressing other factors such as self-shading of the geometry, viewing angle etc. to understand the relation of geometry with energy efficiency in more depth. Similarly, more models of varying shape, floor area and volume should be included in-order to make the surrogate model more useful for all varieties of office buildings and also improve the accuracy of surrogate models by providing more data. In the near future refined tools such as presented in this paper can make a paradigm shift in the design process.

Chapter 3

Using a Surrogate Model to Analyze the Impact of Geometry on the Energy Efficiency of Buildings (Paper: 2)

Bhumika Bhatta[12], Paul Westermann[12], Ralph Evins[12]
[1]Energy in Cities group, Department of Civil Engineering, University of Victoria, B.C, Canada.
[2]Institute for Integrated Energy Systems, University of Victoria, B.C, Canada.

3.1 Abstract

Parametric exploration and optimization of building geometry is a powerful tool for designing energy efficient buildings. However, in practice this process is computationally expensive and time-consuming.

In this research, we explore the use of surrogate models, i.e. efficient statistical approximations of expensive physics-based building simulation models, to lower the computational burden of large-scale building geometry analysis.

For this purpose, we developed a novel dataset of 38,000 residential building models derived from real world floor plans from [69] and train a surrogate model to emulate their simulated annual energy performance. We extract up to 20 parameters as sur-

rogate model inputs to represent the building geometry and show that the trained surrogate model reaches a high accuracy (R^2 score = 0.999, MSE = 0.007 and RMSE = 0.022) on test data. The current setup forms the basis for further research where the complexity of the building models will be increased.

3.3 Key Innovations

- This research provides a framework to develop surrogate models that generalize over any geometry. The framework includes the conversion of 2D floor plan images into 3D building simulation models, the extraction of geometry features, surrogate model training, and performance testing.
- Our first results suggest that off-the-shelf neural network surrogate models provided with engineered geometry features can accurately emulate simulation outcomes. This implies that attempts to use end-to-end deep learning models applied directly to 3D data may be overkill.

3.4 Practical Implications

The motivation of this work is to help architects in integrating energy performance information into early design stage. At this stage, exploration and drastic changes of the building geometry are common. Traditional building simulation approaches cannot keep up with this. We provide surrogate models that estimate the building performance in real time while the building designer is exploring various geometries without the burden of simulating each individual design concept.

We invite other researchers to use our novel dataset containing 38,000 single-zone residential EnergyPlus models of various geometries.

3.5 Introduction

Buildings cover a major share of the global energy consumption and have a large impact on the environment [37].The energy performance of a building is driven by a multitude of impact factors including its shape, location, property, material and others [25]. Among these properties the building geometry hugely affects the energy performance [30] and this, in comparison to other building properties it usually remains constant throughout the life cycle of the building [27].

The early design phase, in the building shape is commonly determined, is crucial for the energy performance of new buildings [48]. However, tools which assess and quantify the impact of geometry at the very early stage of building design are lacking.

In the field of building information modelling (BIM), major research are being undertaken to combine BIM and building simulation to better integrate architectural and building simulation processes [14]. However, building a BIM model can be a major, time-intensive step. Hence, detailed exploration of geometry at the early design phase remains an unconventional, expensive, and time-consuming procedure.

In this work we propose the use of surrogate model driven tools to guide early design decision making. Surrogate models are promising to provide building performance assessment much faster than simulation-based design analysis with low computational cost which overcomes the computational barrier to building performance simulation ([66]).

Many research works have analysed building geometry and energy performance, but some of them use hypothetical shapes ([39]), some limit the geometric features to the shape of facades and do not include footprints ([5]). Similarly work in ([50])focuses only on relative compactness or building volume and energy load.

Research work presented in ([33]) uses a 3D convolutional neural network for the estimation of annual radiation intensities on building facades with a very high accuracy. Similarly work presented in ([57]) shows that high accuracy can be obtained by using deep learning models.

In this research work, we generated a novel dataset, which contains IDF files [20] for 38,000 real world single-zone residential buildings of different geometry. We defined 20 building features to represent building geometry (Table 3.1). Energy data generated from the simulation of generated models along with geometric features are used to train a neural network based machine learning model which approximates the annual energy use intensity (EUI) thus acting as a surrogate model for the detailed simulation. This allows shape to be accounted for the optimization of building features and for design space exploration without the burden of simulating and optimizing building designs individually.

3.6 Methodology

The workflow of this research is shown in the Fig 3.1. To create our novel dataset , we used floor plans from the RPLAN dataset. The RPLAN-dataset is a dataset of 48,000 residential floor plans from real buildings ([69]). The first step of creating our data set was converting the raster images of residential 2D floor plans from RPLAN data set into Vector images then to 3D model and finally into an IDF file ([20]) in the Rhino/Grasshopper environment. The generated building IDF files for the single zoned residential buildings are then simulated in Energy Plus [20]) to get the energy consumption data i.e., heating load, cooling load, total energy use (TEU: Sum of heating, cooling, lighting, appliances, or electricity and gas energy used) and energy use intensity (TEU/ Floor Area) of each building. The energy data was then combined with the geometric features of each building, which was generated from 2D floor plan, in Rhino/Grasshopper. All the models uses only one set of building envelope properties and they are as follows: Wall assembly detail: ”ASHRAE 90.1-2004 Extwall mass climate zone alt-res 5-6” with Nominal U-value of 0.511 [W/m^2k], window detail: double pane window with 2.720 [W/m^2k] and SHGC = 0.761, roof assembly : ASHRAE 90.1-2004 Extroof iead climatezone 5-6 with nominal U-value of 0.351 [W/m^2k] and floor with U-value of 1.072 [W/m^2k].

The final dataset contained geometric data of all 38,000 building and the corresponding energy use data. This was used to train a machine learning model and to investigate the impact of each geometric feature on the energy use.

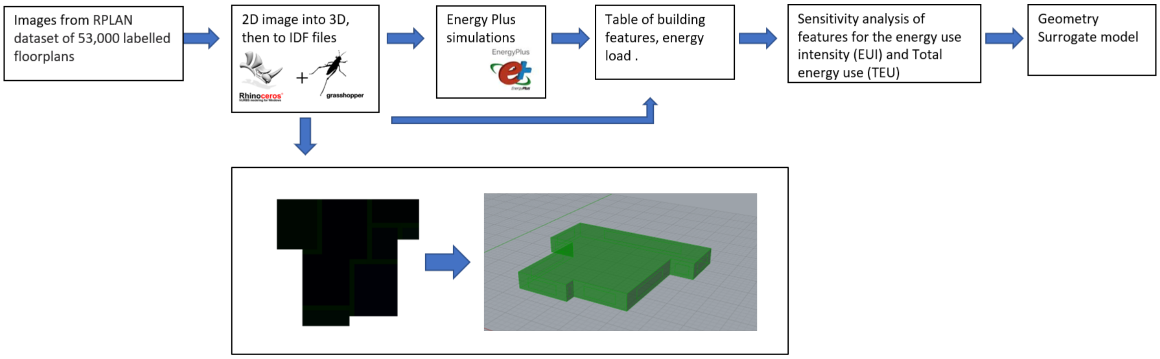


Figure 3.1: Flow-chart showing the workflow of research.

3.6.1 Methodology for creating of our novel dataset.

One of the key components of this research is the use of real-world floor plan of 48,000 residential building. We were able to generate 38,000 IDF files of these geometries. We took 48,000 floor plans from RPlan and attempted to process the images using our workflow. 38,000 images were successfully converted into IDF representations. The remaining 10,000 images had artifacts that made conversion difficult (for example tiny projections). Image cleaning could be implemented to recover these images, but has not been done so far because 38,000 buildings is still enough for our purposes.

The energy data obtained from the simulation of energy models is used to train a surrogate model and study the impact of various aspects of geometry on the energy consumption in the residential building. 2D images of residential floor plans had to be first converted into IDF files in Rhino-Grasshopper [53]. A workflow in Grasshopper (Fig **??**) has been created which takes the raster image of a floor plan converts it into a vector form, scales the image, and adds a vertical plane as a wall surfaces and a horizontal plane as a roof surface and create a 3D model. The next step was to convert the 3D model into IDF format with a plugin in Grasshopper called Ladybug/Honeybee ([53]). By defining a thermal zone in the model along with floor, roof and walls. (Fig. 3.2) shows the assignment of floor, wall, roof ext in the model. The next step was to add the windows. For this experiment we choose the window to wall ratio to be 35% on all facades. The next step was to assign the schedule, and materials to the building. Finally, the model was checked that every element of the building is assigned according to (Fig. 3.2) and exported in IDF format. Geometric features was also collected in the Rhino-Grasshopper environment. This provided the features for the surrogate modelling.

This workflow processes one image at a time. Grasshopper has the flexibility to add new functions, which was used to automate the process (Fig. 3.3). It grabs an image from the specified folder, runs the process above, and when the final step is completed, it clears the window and grabs a new image. With this process we got 38,000 IDF files of different geometries for single family residential building along with a list of parameters describing the building geometry.

3.6.2 Surrogate modelling procedure:

Surrogate models can replace detailed simulations with simplified approximations, thereby sacrificing some accuracy for reduced computational time. We explain the procedure in (Fig. 3.4). It consists of:

- Defining the design problem, i.e., parameters and objectives
- Running a large set of samples in Energy Plus
- Fitting a surrogate model using TensorFlow.
- Analysing the performance of the model using three error metrices: R^2 score, MSE and RMSE ([13]).

Input for the surrogate model is a list of 20 features defining the geometry as shown in Table 3.1, and the output is the energy use intensity (EUI).

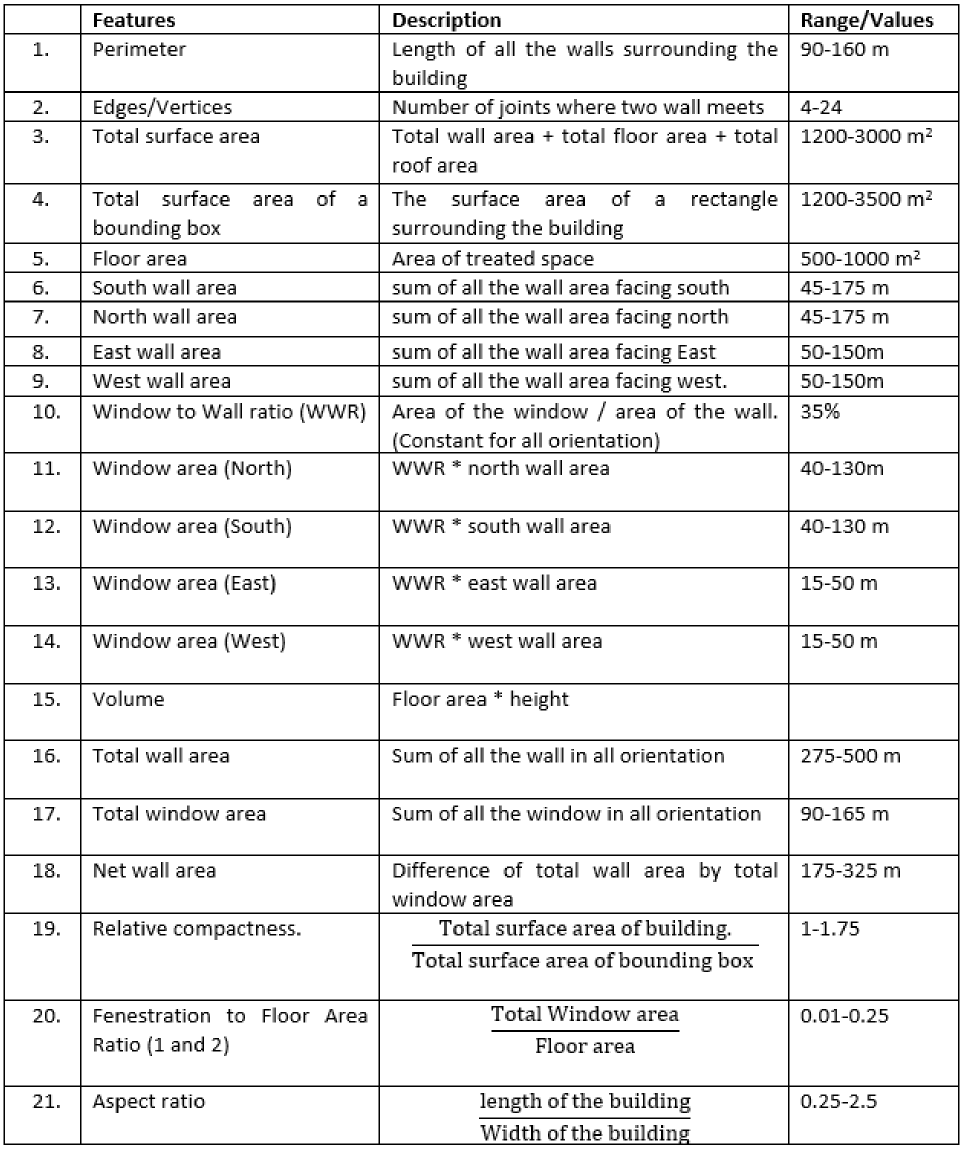
To train the surrogate model, we use the machine learning toolbox Keras, which uses Tensor-Flow as a backend. Tensor-Flow is a toolbox specifically designed to train neural networks. We fit a 1-layered feed-forward neural network on the simulation results and optimize the hyperparameters in a simple grid search (l2-regularization coefficient α, number of neurons per layer $n_{neurons}$). The final surrogate model is fitted using 30,000 simulation samples (plus 7,600 more for testing). The optimal 1-layered network architecture had 21 neurons, where the weights are l2-regularized with $\alpha = 10^{-1}$

3.7 Results of the surrogate model

We analysed the performance of the surrogate model using three different error metrics: R^2 score, MSE and RMSE.

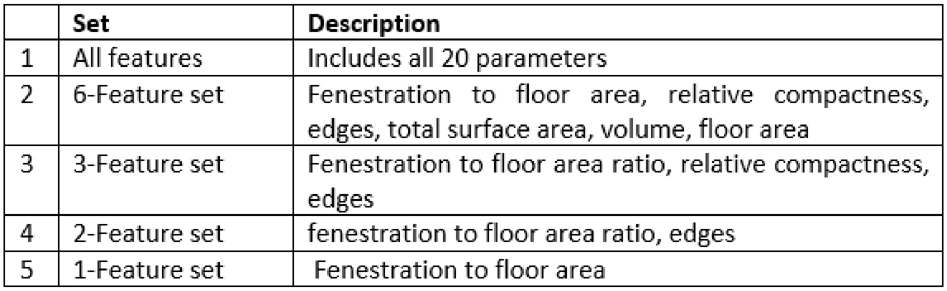
- R^2 score is a statistical way to measure how close the data are to the fitted regression line. R^2 scores range between 0 and 100% where 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variance.
- The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. The smaller the Mean Squared Error, the closer the fit is to the data. The MSE has the units squared of whatever is plotted on the vertical axis.
- The Root Mean Squared Error (RMSE) is the square root of the mean square error. The RMSE is thus the distance, on average, of a data point from the

Table 3.1: Description of building features



	Features	Description	Range/Values
1.	Perimeter	Length of all the walls surrounding the building	90-160 m
2.	Edges/Vertices	Number of joints where two wall meets	4-24
3.	Total surface area	Total wall area + total floor area + total roof area	1200-3000 m^2
4.	Total surface area of a bounding box	The surface area of a rectangle surrounding the building	1200-3500 m^2
5.	Floor area	Area of treated space	500-1000 m^2
6.	South wall area	sum of all the wall area facing south	45-175 m
7.	North wall area	sum of all the wall area facing north	45-175 m
8.	East wall area	sum of all the wall area facing East	50-150m
9.	West wall area	sum of all the wall area facing west.	50-150m
10.	Window to Wall ratio (WWR)	Area of the window / area of the wall. (Constant for all orientation)	35%
11.	Window area (North)	WWR * north wall area	40-130m
12.	Window area (South)	WWR * south wall area	40-130 m
13.	Window area (East)	WWR * east wall area	15-50 m
14.	Window area (West)	WWR * west wall area	15-50 m
15.	Volume	Floor area * height	
16.	Total wall area	Sum of all the wall in all orientation	275-500 m
17.	Total window area	Sum of all the window in all orientation	90-165 m
18.	Net wall area	Difference of total wall area by total window area	175-325 m
19.	Relative compactness.	$\frac{\text{Total surface area of building.}}{\text{Total surface area of bounding box}}$	1-1.75
20.	Fenestration to Floor Area Ratio (1 and 2)	$\frac{\text{Total Window area}}{\text{Floor area}}$	0.01-0.25
21.	Aspect ratio	$\frac{\text{length of the building}}{\text{Width of the building}}$	0.25-2.5

Table 3.2: Cases investigated



	Set	Description
1	All features	Includes all 20 parameters
2	6-Feature set	Fenestration to floor area, relative compactness, edges, total surface area, volume, floor area
3	3-Feature set	Fenestration to floor area ratio, relative compactness, edges
4	2-Feature set	fenestration to floor area ratio, edges
5	1-Feature set	Fenestration to floor area

fitted line, measured along a vertical line ([13]).

The model with 20 features had a high accuracy; (R^2 score = 0.999, MSE = 0.007 and RMSE = 0.022). For the further analysis a correlation heat map in (Fig 3.7), was plotted to understand the importance of each parameter to the output i.e., EUI. From the heatmap, we can see that fenestration to floor area ratio, relative compactness, number of edges, total surface area, volume, and floor area are highly correlated features in comparison to other inputs, although all the other inputs have some co-relation to the output. We created 4 more feature sets (which are the subset of the 20 features set), explained in Table 3.2, with the inputs that had high correlation factor with EUI. With these we trained 4 other surrogate models The accuracy for each feature set is shown in Fig 3.6, R^2 score is between 0.999 to 0.945, MSE score varies from 0.007 to 0.051 and RMSE score varies from 0.021 to 0.2. The results in the Fig 3.6 show that the model with all 20 parameters has high accuracy, in comparison to just the features with high correlation factor suggested by heatmap (Fig 3.7). From the Fig 3.5, that as expected we can see the best performing model is the 20 features set and the worst performing model has just one feature, fenestration to floor area ratio. However, the decrease from 20 features to 6 features is not that bad (R^2 score from 0.999 to 0.984, MSE from 0.007 to 0.015 and RMSE from 0.021 to 0.125.)

3.8 Correlation of parameters and outputs

To understand the nature of relationship between input and output features, we plotted correlation heat map (Fig 3.7) and scatter diagrams for each input with EUI, heating load, cooling load, and total energy use (TEU) (Fig 3.8). The vertical axis are the energy metrics and horizontal are the inputs. This gave more information about the effect of input features with not only EUI but more energy outputs. From the heatmap we see that the features that are most influential for EUI are fenestration to floor area, total surface area, volume, floor area, total wall area, total window area, window areas and number of edges. In Fig (3.8), we see a straight line for TEU with total surface area, volume, and floor area, meaning these features are directly proportional. Similarly, the figure suggests that TEU is also directly proportional to number of edges, but it is a multivariate problem as there are multiple value of TEU for a single number of edges. We also see an interesting plot for EUI and fenestration to floor area. The relation is directly proportional and linear, but there are two different lines for EUI. We plotted a magnified version of EUI vs fenestration to floor area in Figure 3.9 which shows that the bottom line contains all the buildings with a lower number of edges, meaning that with the increase in fenestration to floor area ratio the EUI increases but we can achieve a low EUI with more windows if we stick to buildings with lower number of edges. Figure 3.8 also suggests that the most influential parameters for heating and cooling load are perimeter, total surface area, number of edges, volume, floor area, wall, and window area and parameters except aspect ratio, relative compactness and fenestration to floor have a linear relationship with heating and cooling load.

3.9 Conclusions

In this paper we provide a framework to develop neural network based surrogate models that predicts the EUI over different geometries. The framework consists of conversion of 2D raster images of floor plans into 3D building simulation models, extraction of 20-building features for geometry, surrogate model training, and performance analysis of the impact of each feature. We developed the dataset consisting of 38,000 building simulation IDF files derived from real-world floor plans ([69]). Our results for the training of the surrogate models show that off-the shelve neural network surrogate models provided with manually engineered features are capable of emulating the simulation outcomes really well (R^2 score = 0.999, MSE = 0.007 and RMSE= 0.022) on test data. Overall, this paper shows that a simple machine learning model can perform very well in predicting energy use for various geometries, which provides architects with a great tool to get building performance estimates in real time while they are exploring various designs without the burden of simulating building designs individually.

3.10 Limitation and outlooks

The current setup forms the basis for further research where the complexities will be increased by adding non-geometric envelope parameters such as U-value of walls and windows in the input features. The models used in this research has constant Window to wall ratio(WWR) of 35% on each facades, for the further research we will be varying the WWR as well. No external shading has been used in the windows, which might be the reason we didn't see any relation of individual window area with the EUI. All the models are limited to have one thermal zone with flat roofs and no HVAC system is used, the models are simulated in ideal air load. The energy data used for the research has been obtained from EnergyPlus simulation, and it provides no method to describe thermal bridges in envelope components hence, we didn't see a very distinct relationship of edges with EUI as expected.

This piece of work is basic process to test the principle, since the current workflow only uses one set of building envelope properties. This could still be useful to architects to explore the impact of geometry on a commonly used set of properties, for example code-minimum specifications. The next step will be to explore how the envelope properties interact with the geometry parameters, and how well the surrogate modelling process can handle these interactions.

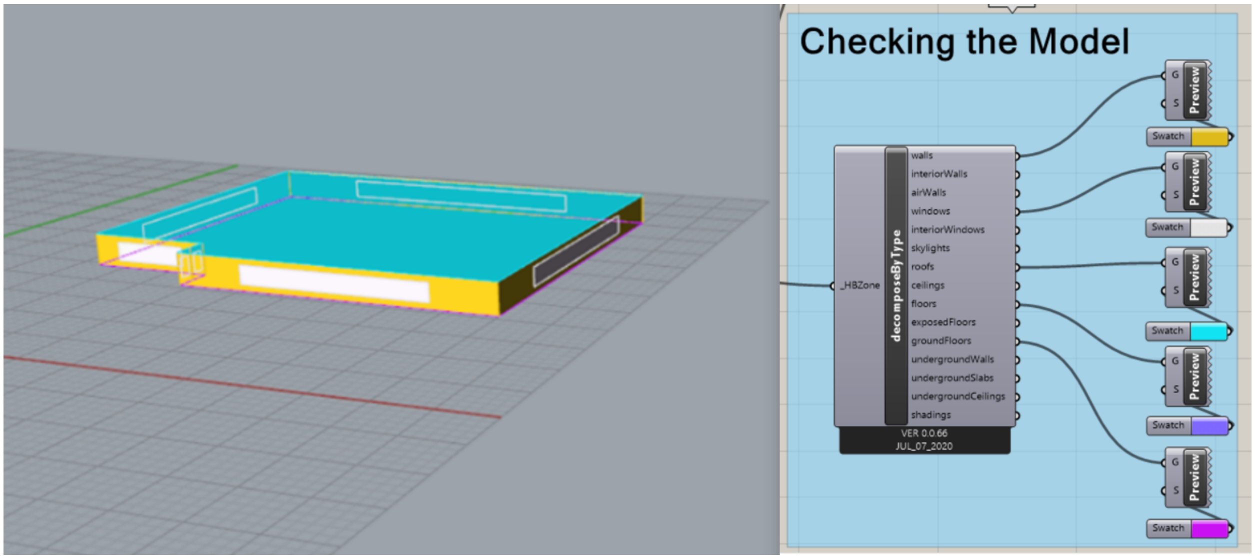
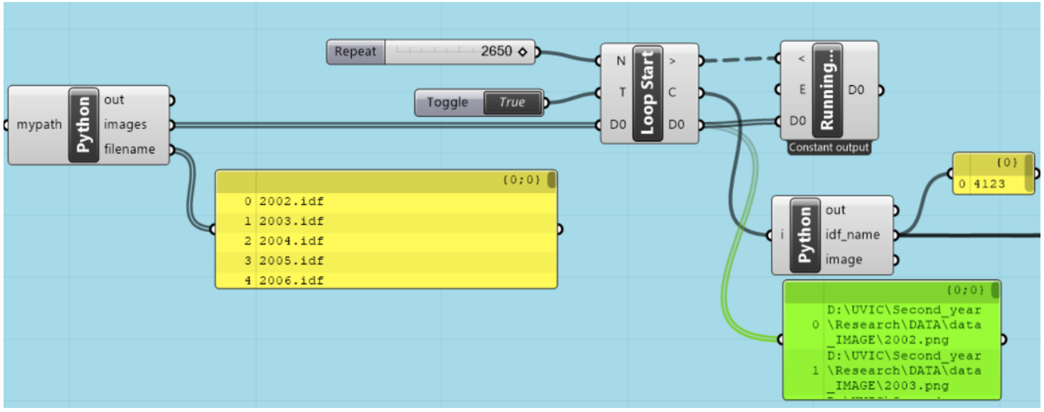


Figure 3.2: Checking the assignment of roof, floor and walls in the created model.



```
{0;0}
0 2002.idf
1 2003.idf
2 2004.idf
3 2005.idf
4 2006.idf
```

```
{0}
0 4123
```

```
{0;0}
  D:\UVIC\Second_year
0 \Research\DATA\data
  _IMAGE\2002.png
  D:\UVIC\Second_year
1 \Research\DATA\data
  _IMAGE\2003.png
```

Figure 3.3: Looping the process for 38,000 images

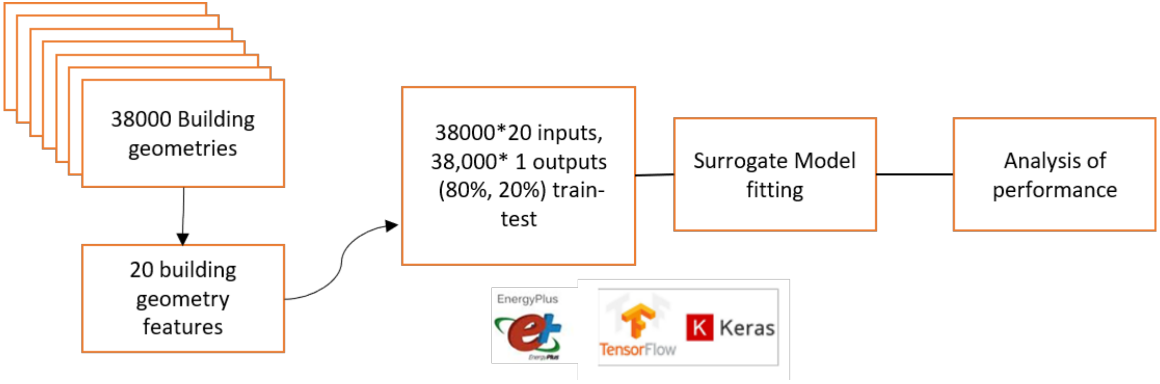


Figure 3.4: Workflow for fitting a surrogate model

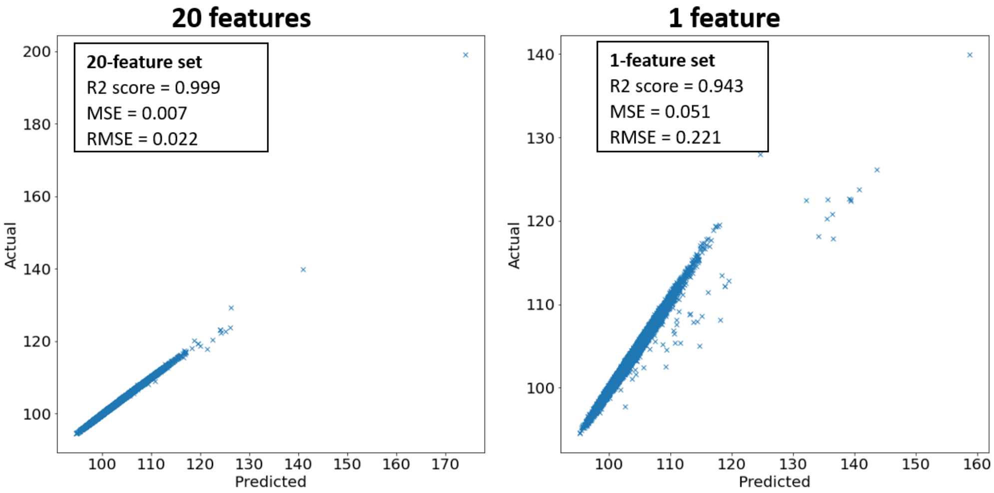


Figure 3.5: Performance of models.

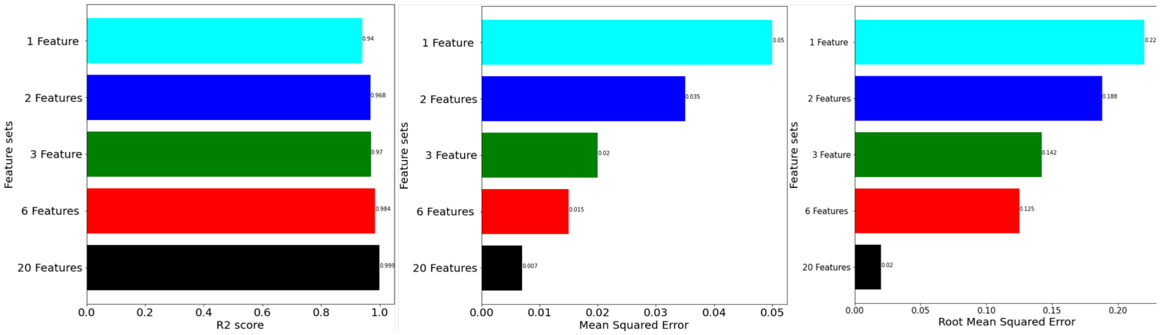


Figure 3.6: Error metrics for each feature set of models.

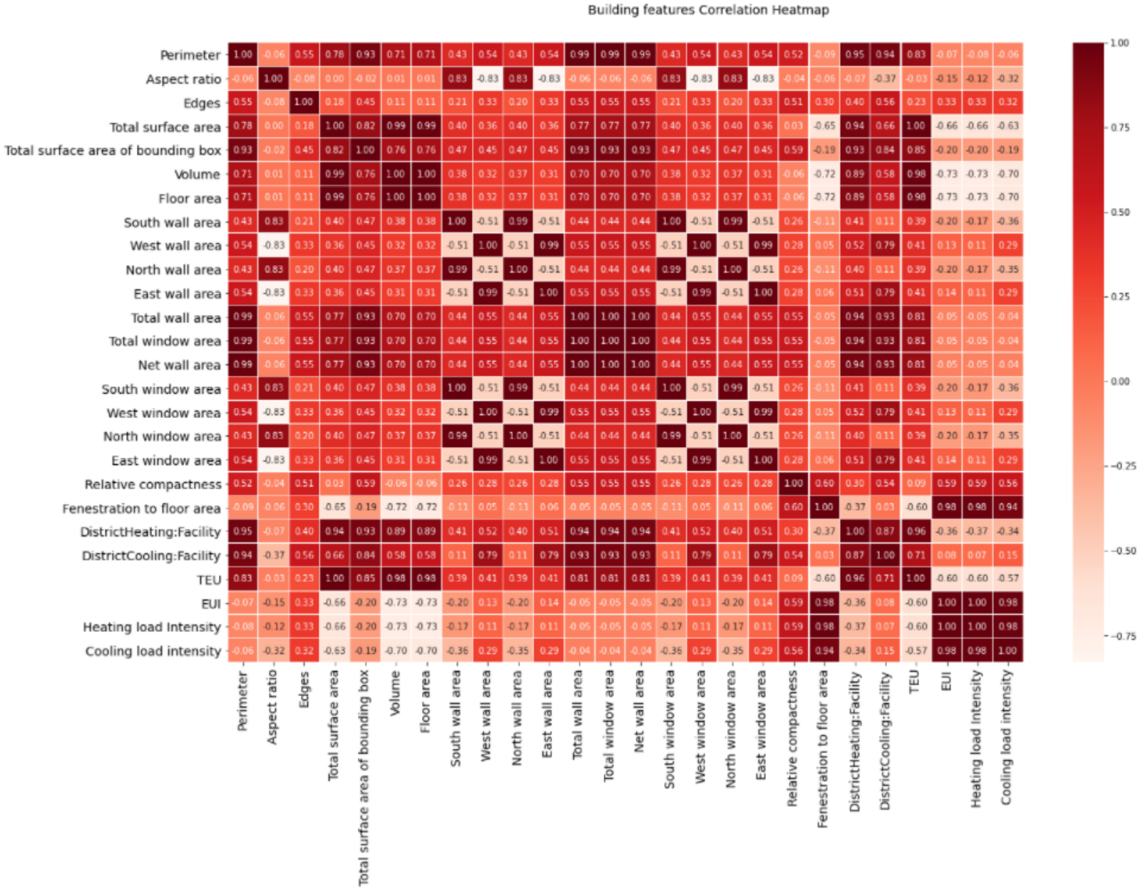


Figure 3.7: Correlation heatmap between inputs and outputs.

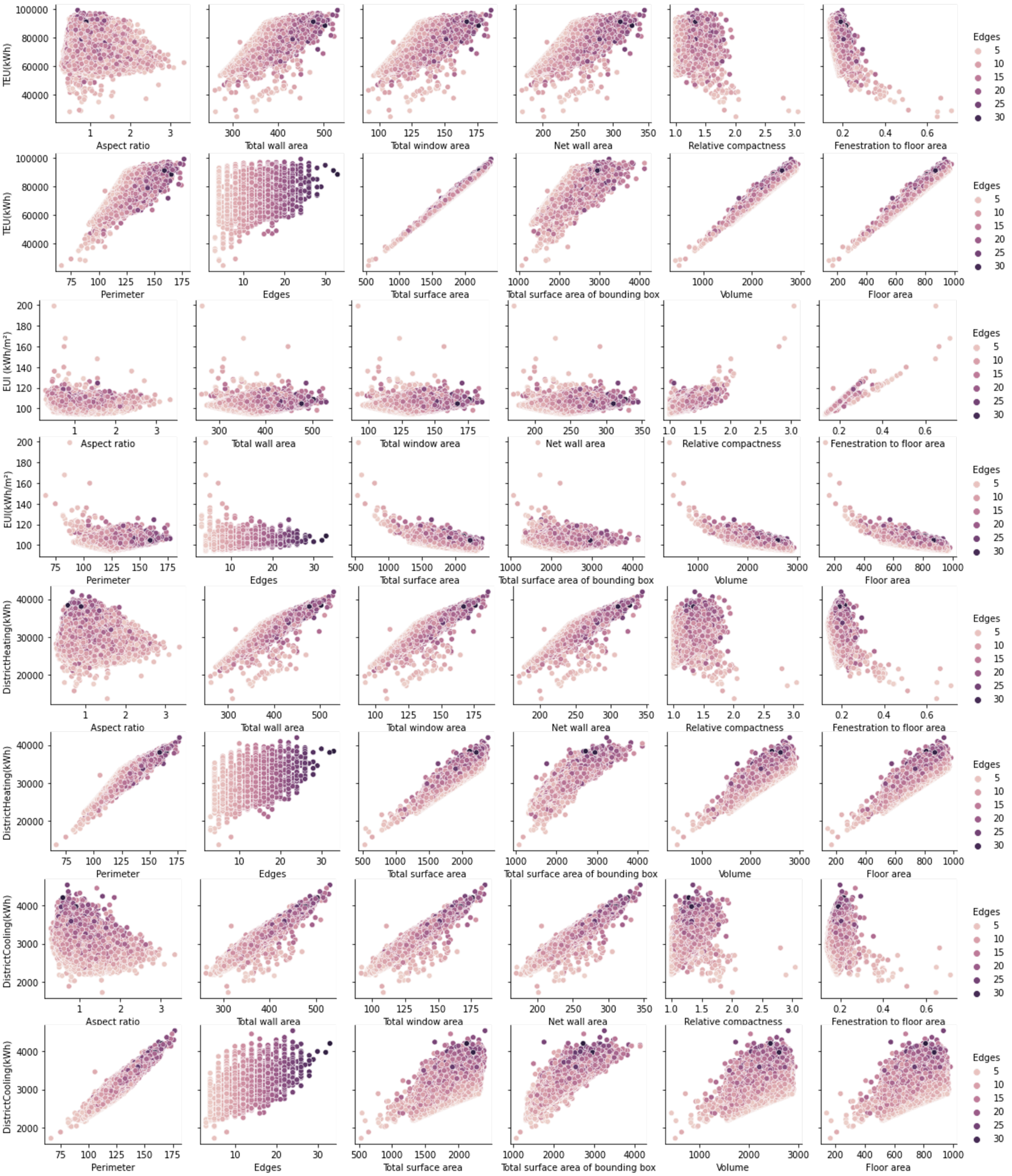


Figure 3.8: Relation of inputs with energy outputs

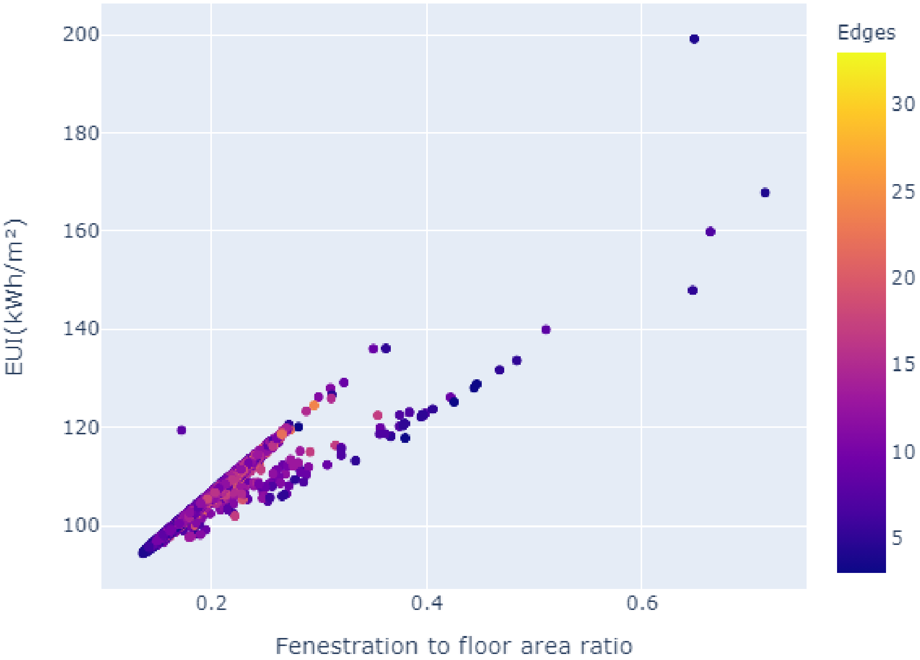


Figure 3.9: EUI versus Fenestration to floor area ratio colored by number of edges.

Chapter 4

Comparison of monolithic and component-based surrogate models to analyze the impact of geometry on the energy efficiency of buildings (Paper:3)

Bhumika Bhatta[1], Paul Westermann[1], Ralph Evins[1]
[1]Energy in Cities group, Department of Civil Engineering, University of Victoria, B.C, Canada.

4.1 Abstract

Parametric exploration and optimization of building geometry is a powerful tool for designing energy-efficient buildings. However, in practice, this process is computationally expensive and time-consuming. In this research, we explore the use of surrogate models, i.e. efficient statistical approximations of expensive physics-based building simulation models, to lower the computational burden of large-scale building geometry analysis. We explored two different modelling techniques, monolithic and component-based surrogate models, and compared their performance. Monolithic models are fitted all in one go, whereas component-based models fit individual

models for different aspects of the design, in this case, envelope and geometry, then combine them together. 4.1 shows the overall workflow of the research work presented in this paper.

We develop 6 different models (2 monolithic and 4 component-based models) with different combinations of inputs. For this purpose, we developed a novel dataset of 38,000 residential building models derived from real-world floor plans from [36] and trained a surrogate model to emulate their simulated annual energy performance. We extract up to 20 geometry parameters and 10 envelope parameters as surrogate model inputs to represent the building geometry and envelope. The work confirms that expected monolithic models perform better, but the performance of the component-based model also falls under an acceptable range of accuracy for a conceptual design phase. The best performing monolithic model had an R^2 score of 0.993 mean squared error (MSE) of 0.006, and root mean squared error (RMSE) of 0.0077; the best performing component-based surrogate model had an R^2 score of 0.978, MSE 0.0203, and RMSE 0.142.

4.2 Key Highlights

1. We develop neural network-based surrogate models for geometry and envelope features to predict energy use in buildings.
2. We compared the performance of component-based and monolithic models.
3. We compared the performance with many different combinations of features.
4. Flexible component-based models are shown to perform nearly as well as monolithic.

4.3 Introduction

Building accounts for a huge share of global energy consumption, with a great impact on the environment [37]. Preventing climate change requires a change in the building industry and this calls for high-performance building designs and operations.

The energy consumption of a building is dependent on many factors including its shape, location, envelope properties, mechanical system, etc [11]. Among this multitude of properties, the shape of the building hugely affects the energy performance

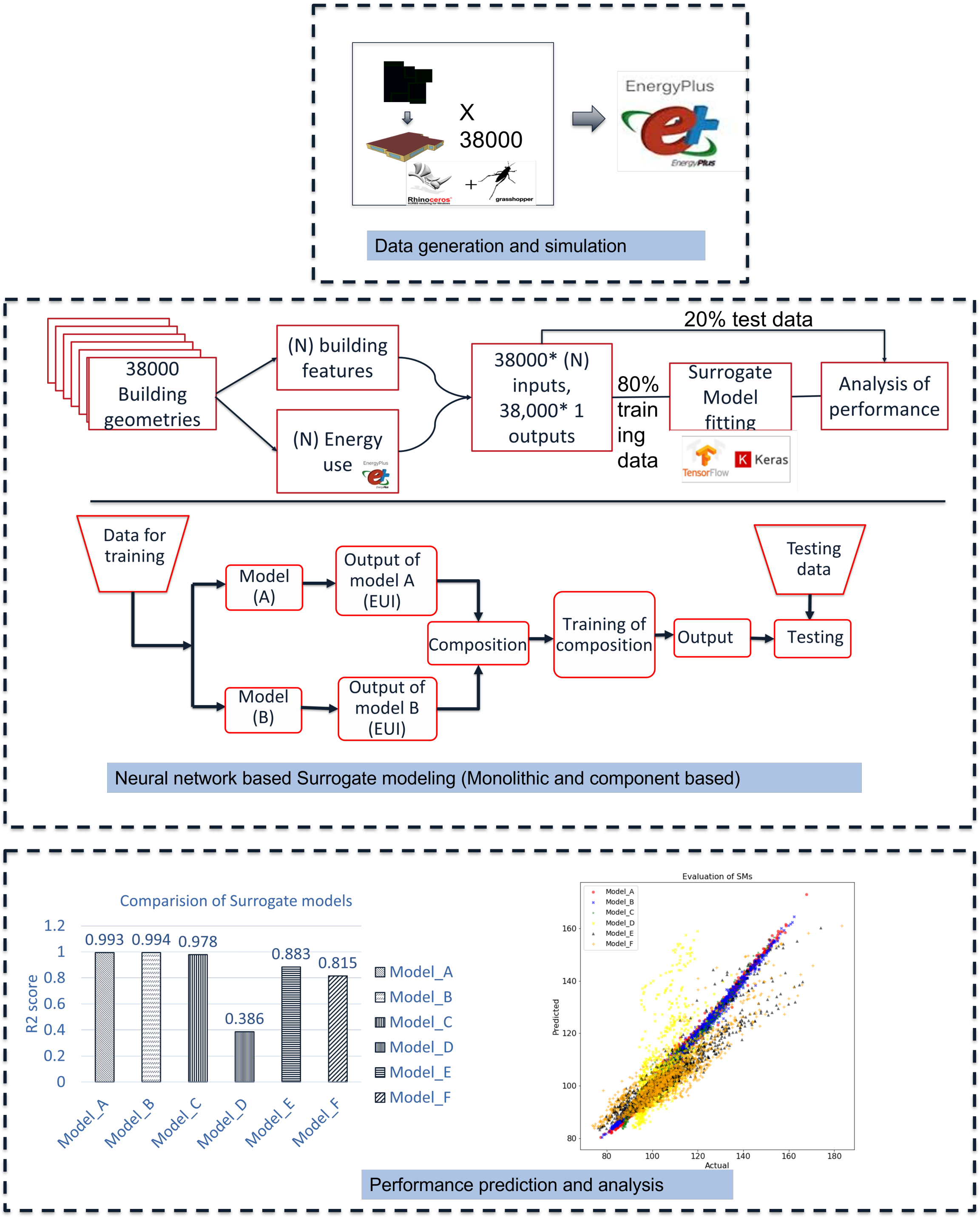


Figure 4.1: Graphical abstract showing the workflow of the research.

of a building [30], [15]. This is a property that usually remains constant throughout the life cycle of a building [27]. The building footprint is usually determined at an early design phase, which is a critical phase for making design decisions for the building [48]. Therefore, it is important to facilitate efficient and sustainable design feedback at this crucial point in the process [62], [70], [10], [65], [8]. However, our industry lacks the tools to assess and quantify the impact of geometry at a conceptual phase making designers rely more on rules of thumb. [70] explains that some existing building simulation tools claim to be useful in the conceptual phase but they provide little assistance for architects/designers until the end of the conceptual design stage when large changes of the design can not be made affordably. Also, many simulation tools such as EnergyPlus [19], Hot2000 [32], etc. were originally developed for engineers so the input/output parameters and the description method of the available information are different from an architect's preferred method of declaring parameters [70]. Nonetheless, designers are keen to use building performance simulation (BPS) information to improve their designs [73]. Traditional BPS software is also time-consuming and the quality of results depends on simplification methods applied [9] and on the skill level of the energy modeler [55].

Surrogate models, also known as meta-models, use machine learning techniques to extract knowledge from the BPS data by recognizing patterns within them [57]. Surrogate models are now used to replace detailed building simulations as it lowers the burden of simulating each design solution at the conceptual phase which are generally time-consuming and computationally expensive [66]. Surrogate models, also known as statistical models, use machine learning to predict performance based on initial simulations within a parametric design space. These models leverage learned relationships from prior computation and replace detailed simulation with simplified approximation, with some accuracy sacrificed but still reasonable for the relative sense of performance required when making conceptual phase design decisions. Research has made it evident that they can be used as an alternative for the common building simulations in many cases [66].

Component-based surrogate models extend the practicality and usability of the surrogate models as different design problems can be explored without rebuilding the whole model. Lastly, we show a sensitivity analysis of the features with the output using a correlation heatmap. The heatmap shows the important geometry and envelope features along with their pattern of a relationship with energy use. The results from the heatmap were used for feature engineering. We only selected

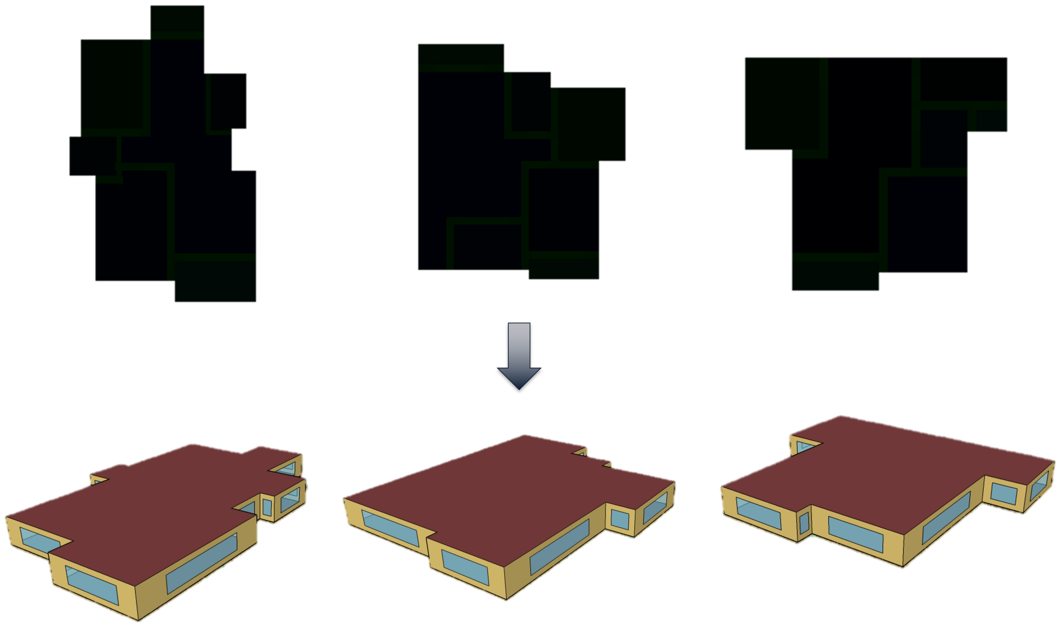


Figure 4.2: Example of conversion of the 2D floor plan into 3D models and to EnergyPlus model using Rhino/Grasshopper environment

the important parameters from the geometry and envelope features sets to train the surrogate models.

This research has the following major objectives:

- We show that surrogate models can capture the impact of geometry on the energy use of the building by using monolithic and component-based neural network-based surrogate models. For the experiments, we use real-world datasets and a list of 30 features to define important geometry and envelope features.
- We show that dividing surrogate models up into components can allow reasonable accuracy compared to monolithic models and this would therefore allow models to be fitted separately for the geometry aspect from the envelope.
- We then compare the performance of the component-based model with the monolithic model and also discuss the flexibility of component-based surrogates to train a model on different building features such as geometry and envelope.
- We analyze the impact of different geometry and envelope features on the energy

use of the building. We also compare the performance of the models developed with different combinations for feature engineering.

In this research, we use 38,000 building models to generate 4 different datasets. We used labeled 2D images of residential floor plans from [36] and generated a novel dataset of 38,000 building models as shown in 4.2. We developed a process to convert 2D images into 3D models and to export EnergyPlus models in a Rhino/Grasshopper environment [12]. We then simulated these building models to generate datasets to train the surrogate models. We derived various subsets with various combinations of geometry and envelope features to train multiple surrogate models to capture the impact of geometry and envelope features. We trained multiple monolithic models and also composed component-based surrogate models with two sub-models (geometry and envelope) and finally compared the performance with each other.

4.4 Literature Review

Many research studies considered geometry for energy efficiency in buildings but they are usually limited to analyzing only a few features such as relative compactness [50], surface area, roof area, wall area [54], volume relative compactness [43], and orientation [49], etc. Similarly, the work presented in [5] limits the geometric feature to the shape of the facade and doesn't account for the impact of footprint. [63] defines 8 common building geometry parameters that affect the energy efficiency of the building. Research presented in [34] used Monte Carlo simulation to predict the energy consumption reflecting the design uncertainty and sensitivity to identify important parameters which included less geometry and envelope in the meta-model output along with heating, ventilation, and air conditioning parameters.

Research has been done for the exploration of design space and optimization of design parameters for energy-efficient design which requires building performance simulation for different design options [18], [40]. Singh et. al [59] developed a plugin for building information modelling to develop building performance simulation and tested it in over five design options for office buildings. However, state-of-the-art dynamic simulation tools used for BPS during the design stage are still computationally expensive and time extensive. The work presented in this paper provides a machine learning model which replaces the detailed BPS and makes the design space exploration for geometry and envelope features less time-consuming and less computationally expensive with some sacrifice in the accuracy making it suitable for early design exploration.

In our research work, we introduce 20 features to define geometry along with 10 features for the envelope. Our datasets contain variations in geometry as well as envelope features. Other researchers such as [39] analyze geometry but based on hypothetical shapes. Martin et al [45] compared the cooling load for single-zone and multi-zone models for just a rectangle shape. Most of the research work uses a hypothetical building and or has limitations in the sample size of geometry whereas the work presented in this paper uses a novel data set of 38,000 residential building models derived from real residential floor plans from [69].

A lot of research has been done to make energy modelling simulations less computationally expensive, less time-consuming, and more insightful for the design process. Initially, simplified analytical models were used. More recently, meta-modelling techniques have been used, including regressions such as SVR etc. [36], [7], [17], [22], neural network [17], [64], [28], classification and regression tree [17], and ensemble models [17], [55], [57], [56]. Among different machine learning techniques, Artificial Neural Networks are the most popular as they are convenient and easy to use by a regular operator after the model has been established [57]. Due to this a lot of studies have been done that have explored the use of neural networks for predicting diverse aspects of energy efficiency and the performance of buildings [71], [11], [24], [47], [35], [22], [41], [38]. However, most works have reported unsatisfactory error rates, lack of generalizability and most have considered only a few parameters that affect the building energy use. Research work presented in [58] used 2D Convolutional Neural Network for shape information along with other building envelope parameters such as U-values of the wall, window, roof, etc to train a surrogate model on Energy Use Intensity prediction. They also trained another model on building parameters such as area, volume as shape information, and other building envelope parameters using Deep Neural Network. The comparison between those two models showed that 2D-CNN performed slightly better but 2D-CNN does not give any information about which geometry parameters were important.

The work presented in [39] explains the framework for the composition of several machine learning models to solve a classification problem. We use a similar technique to solve our regression problems by applying compositional machine learning methods in the context of different building design areas. The model developed gives the user flexibility to independently choose and set their design parameters with the possibility of further re-using the models themselves and the compositions made from them. [55], [57] uses a component-based surrogate modelling technique to predict the heating load and cooling load of the building and evaluate their practicability for design space exploration. However, the geometry inputs used in the paper are limited to length, width, WWR, and orientation whereas in our research work we have a set of 20 geometry features.

In our work, we not just evaluate the performance of component-based surrogate models with BPS but also compare the accuracy with the monolithic models trained on similar data. The data set used in our research has more variation in geometry as compared to [57]. They use 800 design combinations resulting in a 9600 training dataset whereas we used 36,000 geometry resulting in a 1,800,000 training dataset. We also trained our surrogate models with various combinations of datasets simulated in different conditions to best capture the impact of geometry for the energy use of buildings. Training on various datasets helped in understanding the important and relevant features.

4.5 Methodology

We simulated 38,000 building models, to generate the training data containing building features and energy outputs. The datasets we generated have various combinations of geometry and envelope features which we use to train multiple surrogate models to emulate the energy use intensity (EUI) of the building.

Where, EUI is Total energy use (TEU)/ floor area, Total energy use is the sum of electricity, gas, heating, and cooling load.

Figure 3 shows the different combinations of surrogate models investigated. We simulate four different data sets with our 38,000 building models and further derive their subsets to train different models. Figure 4 explains the four dataset and their subsets that were used to train our 10 different surrogate models. Model A is trained on subset A (common data A) and Common surrogate Model B is trained on subset B (Common data B), both are monolithic surrogate models that predict the EUI. Model C is a composition of two separately trained surrogate models. The two monolithic surrogate models are geometry $surrogate_c$ and envelope $surrogate_c$. Both surrogate models are trained on subsets of a common dataset. In this model, only the envelope dataset has window information in the form of WWR.

For the workflow shown in E and F of 4.3, we use two different training sets, sim-

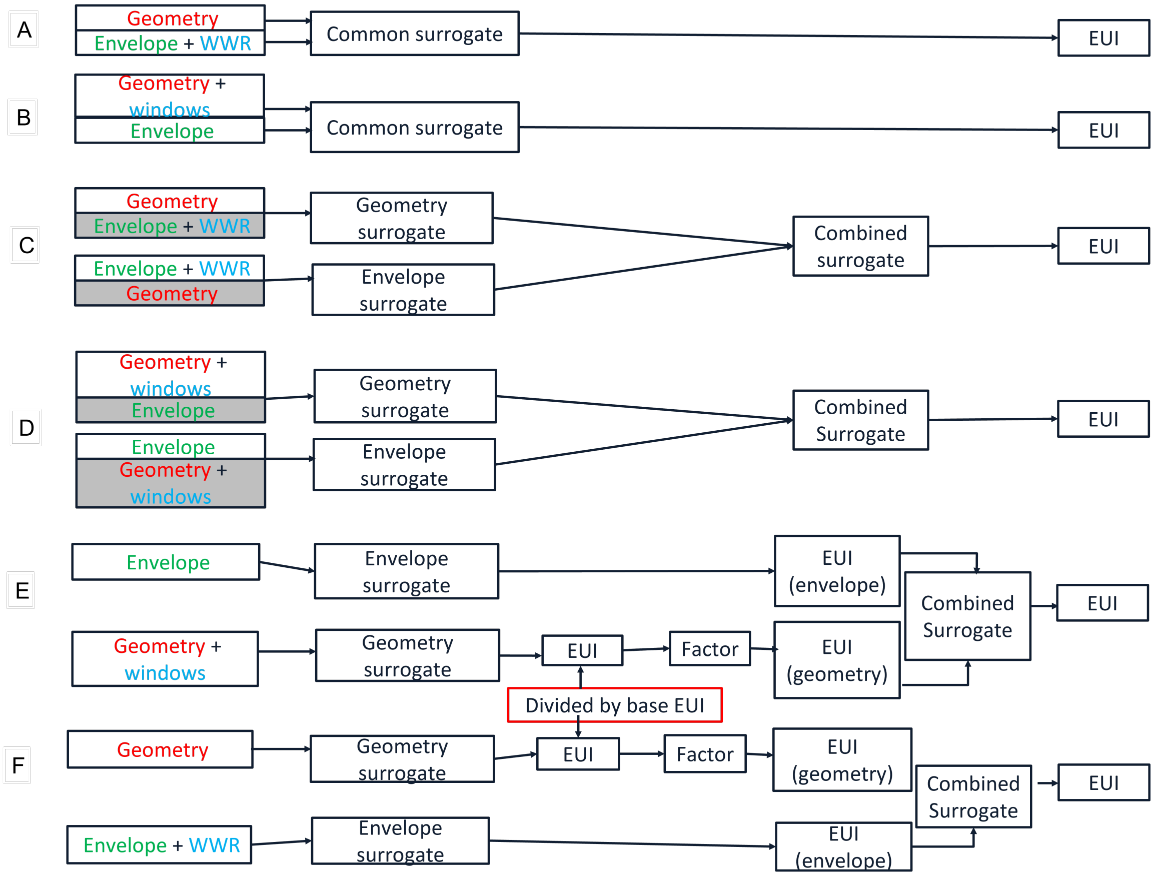


Figure 4.3: Flow chart describing the different surrogate model combinations investigated.

ulated in different conditions to train the base models (Geometry surrogate and envelope surrogate). Hence to normalize these models, we add an additional step where we divide the output of the geometry surrogate by basing the EUI value. The base EUI value (100 KWh/m^2 is obtained by stimulating a base geometry model (Square shaped with 35% WWR on each side) with base envelope features ("ASHRAE 90.1-2004 Extwall mass climate zone alt-res 5-6" with Nominal U-value of 0.511 [W/m^2k], window detail: double pane window with 2.720 [W/m^2k] and SHGC = 0.761, roof assembly : ASHRAE 90.1-2004 Extroof iead climate zone 5-6 with nominal U-value of 0.351 [W/m^2k] and floor with U-value of 1.072 [W/m^2k]). We then multiply the geometry factor with the EUI obtained from the envelope surrogate model to get the final EUI. Model A, B, C, and D use the same common dataset but have a unique set of input and hidden parameters. The dataset for geometry submodel for E and F are the same but they also have a unique set of input and hidden parameters. The dataset for envelope submodel for model E was simulated with constant WWR whereas the dataset for the envelope submodel of model F was generated by simulating variation in WWR.

Combined surrogate model D is also a composition of two separately trained monolithic surrogate models. The two surrogate sub-models are geometry surrogate and envelope surrogate. Both surrogate models are trained on subsets of a common dataset. In this model, only the geometry dataset has window information in the form of window areas and a sum of window areas.

After training the models in a common data set we also simulated geometry data and envelope data independently. Combined surrogate model E and model F are also the composition of geometry and envelope surrogate which were trained on different datasets (independently simulated geometry data and envelope data). Geometry sub-model e and geometry sub-model f are trained on a dataset obtained by simulating various geometries with constant envelope features (Base envelope features). The only difference between geometry sub-model e and geometry sub-model f is that sub-model e has window information and sub-model f does not. Similarly, the envelope models used for combined model E and model F have separate datasets. Envelope sub-model e uses a dataset obtained by simulating base IDF geometry with 6 envelope features (not including WWRs) whereas envelope sub-model f is trained on envelope dataset obtained by simulating base IDF geometry with 10 features (Including WWRs).

- Model A was trained on a common dataset with windows detail as WWR in envelope parameters. This is a monolithic neural network-based surrogate model. Model A has window details in only envelope features set. It has a total of 17 input parameters and 9 hidden parameters. Here the hidden parameters are the window detail in the geometry feature set as shown in “A” figure 4.4

- Model A was trained on a common dataset with windows detail as WWR in envelope parameters. This is a monolithic neural network-based surrogate model. Model A has window details in only envelope features set. It has a total of 17 input parameters and 9 hidden parameters. Here the hidden parameters are the window detail in the geometry feature set as shown in “A” figure 4.4

- Model B was also trained on a common dataset with window detail as window areas in geometry parameters. This is a single neural network-based surrogate model. Model B has window details in the geometry features set. It has a total of 22 input parameters and 4 hidden parameters. Here the hidden parameters are the window to wall ratios per orientation in the envelope feature set as shown in “B” of figure 4.4.

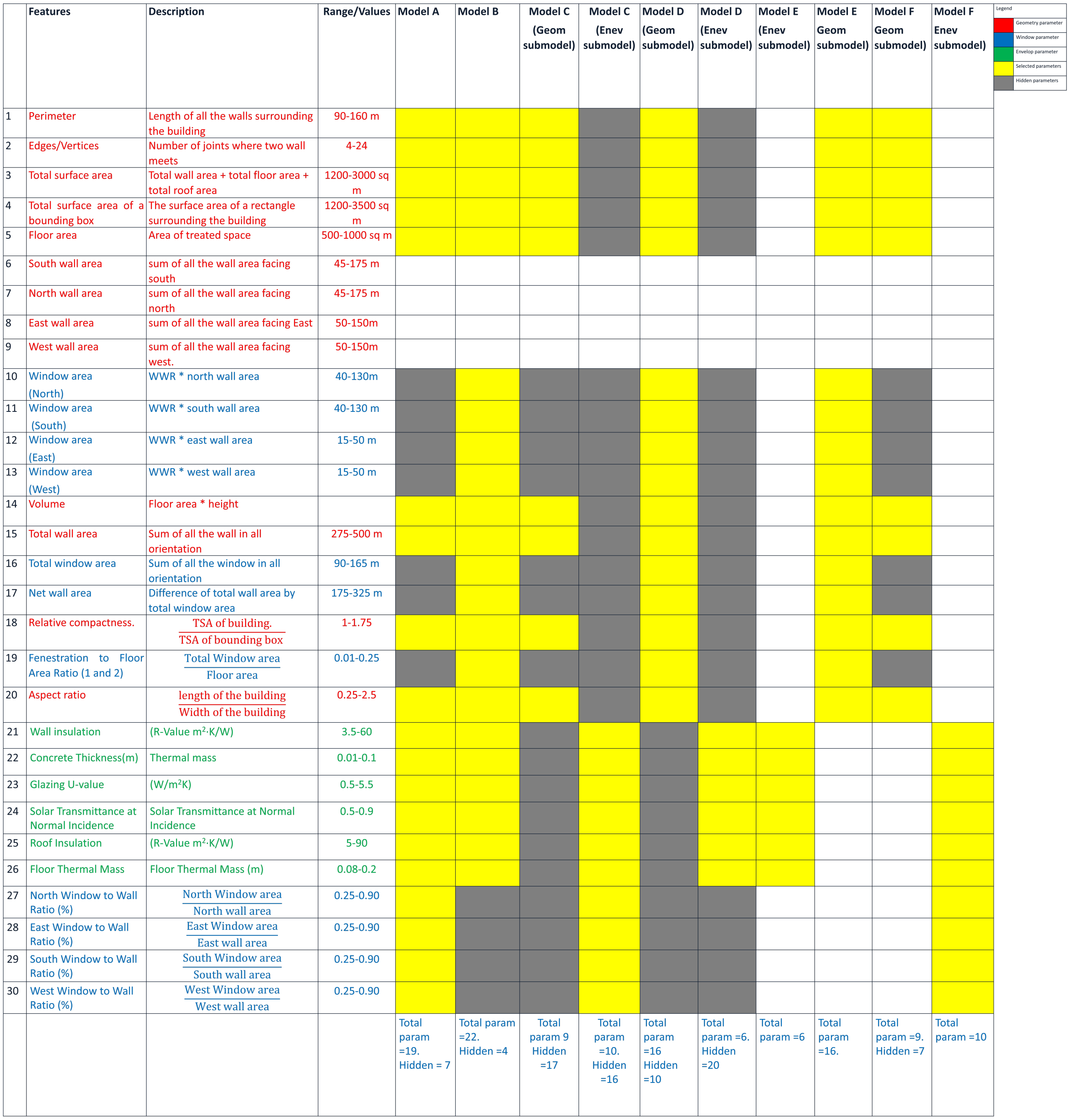
4.5.1 Generation of Datasets

The first data set (also known as a common data set) has a combination of geometric as well as envelope features of the building along with its energy use intensity. This dataset contained 38,0000 unique floor plans (varied geometry) along with variations in their WWR and envelope properties. Each variation represented a unique building, which was then simulated to obtain their energy use intensity from EnergyPlus for Victoria weather. Surrogate models A and B are trained with the subset of the common data set and have information on both geometric and envelope aspects of the building as shown in 4.4. The only difference between training data of Surrogate model A and B is that the information of the window in model A is in terms of a window to wall ratio whereas model B gets the window details from each window area. Theoretically, both the models have the same information but in a different format and feature type. We then split the features from the common dataset into geometry features and envelope features as shown in 4.4.

The second dataset type is the geometry dataset. The geometry data set was obtained by simulating 38,000 unique floor plans with a constant value of envelope features and a constant window to wall ratio of 35% for each building. The building features in the geometry data set are shown in Figure 10. The variation in window area for each building in this dataset is due to variation in wall area which is a result of the unique floor plans of each building. This data set was used to train a geometry surrogate model which was then used to make a combined surrogate model E and F.

The third dataset is an envelope dataset without WWR. The envelope data set without WWR was obtained by simulating a base IDF (square-shaped) with various combinations of 6 envelope features(excluding window to wall ratio per orientation). This dataset has no window detail.

The fourth dataset is an envelope dataset with WWR, which was obtained by simulating a base IDF (square-shaped) with various combinations of 10 envelope features (Including window to wall ratios). The parameters for the envelope data set are shown in 4.9. We simulated 10,000 combinations of envelope features. This data set was used to train an envelope surrogate model used in Model E and F respectively.



	Features	Description	Range/Values	Model A	Model B	Model C (Geom submodel)	Model C (Enev submodel)	Model D (Geom submodel)	Model D (Enev submodel)	Model E (Enev submodel)	Model E Geom submodel)	Model F Geom submodel)	Model F Enev submodel)
1	Perimeter	Length of all the walls surrounding the building	90-160 m	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
2	Edges/Vertices	Number of joints where two wall meets	4-24	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
3	Total surface area	Total wall area + total floor area + total roof area	1200-3000 sq m	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
4	Total surface area of a bounding box	The surface area of a rectangle surrounding the building	1200-3500 sq m	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
5	Floor area	Area of treated space	500-1000 sq m	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
6	South wall area	sum of all the wall area facing south	45-175 m										
7	North wall area	sum of all the wall area facing north	45-175 m										
8	East wall area	sum of all the wall area facing East	50-150m										
9	West wall area	sum of all the wall area facing west.	50-150m										
10	Window area (North)	WWR * north wall area	40-130m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
11	Window area (South)	WWR * south wall area	40-130 m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
12	Window area (East)	WWR * east wall area	15-50 m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
13	Window area (West)	WWR * west wall area	15-50 m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
14	Volume	Floor area * height		🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
15	Total wall area	Sum of all the wall in all orientation	275-500 m	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
16	Total window area	Sum of all the window in all orientation	90-165 m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
17	Net wall area	Difference of total wall area by total window area	175-325 m	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
18	Relative compactness.	$\frac{\text{TSA of building.}}{\text{TSA of bounding box}}$	1-1.75	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
19	Fenestration to Floor Area Ratio (1 and 2)	$\frac{\text{Total Window area}}{\text{Floor area}}$	0.01-0.25	⬛	🟨	⬛	⬛	🟨	⬛		🟨	⬛	
20	Aspect ratio	$\frac{\text{length of the building}}{\text{Width of the building}}$	0.25-2.5	🟨	🟨	🟨	⬛	🟨	⬛		🟨	🟨	
21	Wall insulation	(R-Value $m^2 \cdot K/W$)	3.5-60	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
22	Concrete Thickness(m)	Thermal mass	0.01-0.1	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
23	Glazing U-value	(W/m^2K)	0.5-5.5	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
24	Solar Transmittance at Normal Incidence	Solar Transmittance at Normal Incidence	0.5-0.9	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
25	Roof Insulation	(R-Value $m^2 \cdot K/W$)	5-90	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
26	Floor Thermal Mass	Floor Thermal Mass (m)	0.08-0.2	🟨	🟨	⬛	🟨	⬛	🟨	🟨			🟨
27	North Window to Wall Ratio (%)	$\frac{\text{North Window area}}{\text{North wall area}}$	0.25-0.90	🟨	⬛	⬛	🟨	⬛	⬛				🟨
28	East Window to Wall Ratio (%)	$\frac{\text{East Window area}}{\text{East wall area}}$	0.25-0.90	🟨	⬛	⬛	🟨	⬛	⬛				🟨
29	South Window to Wall Ratio (%)	$\frac{\text{South Window area}}{\text{South wall area}}$	0.25-0.90	🟨	⬛	⬛	🟨	⬛	⬛				🟨
30	West Window to Wall Ratio (%)	$\frac{\text{West Window area}}{\text{West wall area}}$	0.25-0.90	🟨	⬛	⬛	🟨	⬛	⬛				🟨
				Total param =19. Hidden = 7	Total param =22. Hidden =4	Total param 9 Hidden =17	Total param =10. Hidden =16	Total param =16 Hidden =10	Total param =6. Hidden =20	Total param =6	Total param =16.	Total param =9. Hidden =7	Total param =10

Legend	
🟥	Geometry parameter
🟦	Window parameter
🟩	Envelop parameter
🟨	Selected parameters
⬛	Hidden parameters

Figure 4.4: Features for the surrogate model with their description.

4.5.2 Surrogate Modelling Process

The modelling procedure for fitting a surrogate model is shown in figure 4.5. It consists of

1. Data generation by running a large set of samples in EnergyPlus and model type selection.
2. Setting the design problem i.e design parameters and objectives.
3. Fitting the surrogate model using Tensorflow and Keras.
4. Performance analysis using error metrics such as the R^2 score, MSE, and RMSE [13]

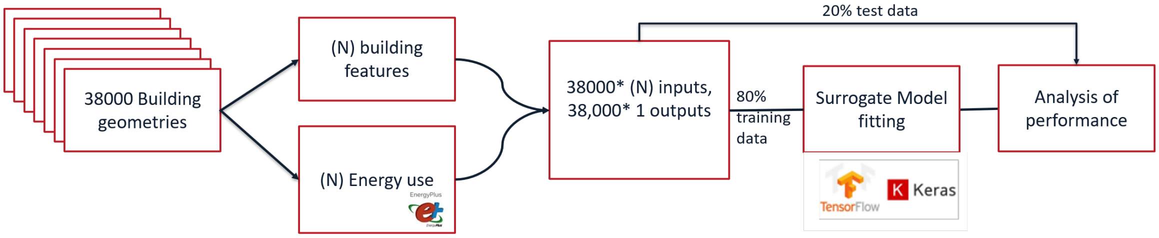


Figure 4.5: Flowchart showing the procedure of fitting a surrogate model

To train the surrogate model, we use the machine learning toolbox Keras, which is built on TensorFlow as an advanced neural network library developed by Google [45]. We fit a 1-layered feed-forward neural network on the simulation results and optimize the hyperparameters in a simple grid search (l2-regularization coefficient , number of neurons per layer n_n). The final surrogate model is fitted using 30,000 simulation samples (with 7,600 more used for testing). The optimal 1-layered network architecture had neurons equal to N (number of features) *1, where the weights are l2-regularized with $\alpha = 10^{-1}$. We analyzed the performance of the surrogate model using three different error metrics: R^2 score, MSE, and RMSE [13].

R^2 score is a statistical way to measure how close the data are to the fitted regression line. R^2 scores range between 0 and 100% where 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variance.

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. The smaller the Mean Squared Error, the closer the fit is to the data. The MSE has the units squared of whatever is plotted on the vertical axis.

The Root Mean Squared Error (RMSE) is the square root of the mean square error. The RMSE is thus the distance, on average, of a data point from the fitted line, measured along a vertical line.

4.5.3 Component Based Surrogate Modelling

To analyze the energy efficiency in building designers rely on information such as heat flow, internal gains, heat gain via windows, etc and there is a multitude of factors that affect this information which will ultimately help the designers to make the right design decisions. But all the necessary information is not present in a monolithic model. [57],[56]. For example, a surrogate model trained to emulate the energy use of a building only based on envelope features will not be able to perform well when the HVAC features are introduced in the input. To increase the usability of a monolithic model, would require data collection from various possible design options and a complex model architecture [55]. Figure 4.6 shows the process of training a component-based surrogate model.

This research work consists of two surrogate model training techniques. The first one consists of monolithic models obtained by training a model with a data set containing geometry and envelope features at once. Model A and Model B are developed with this technique. The second modelling technique is a component-based surrogate model, trained with the composition of two different surrogate models. In this method, two base surrogate models are trained separately after which we take the output of each model as an input for a third combined surrogate model. This method is used for Model C-F.

In this work, we develop component-based surrogate models by combining two different monolithic sub-models for geometry and envelope features. We have 6 different final surrogate models of which four are component-based surrogate models trained by the composition of two different models for geometry and envelope. We use the emulated output (EUI) from both models and use that as input for the combined model. The combined model will then have 2 inputs (Individual EUI) and one output (Final EUI). The combined model is then tested with a common testing dataset for the performance analysis.

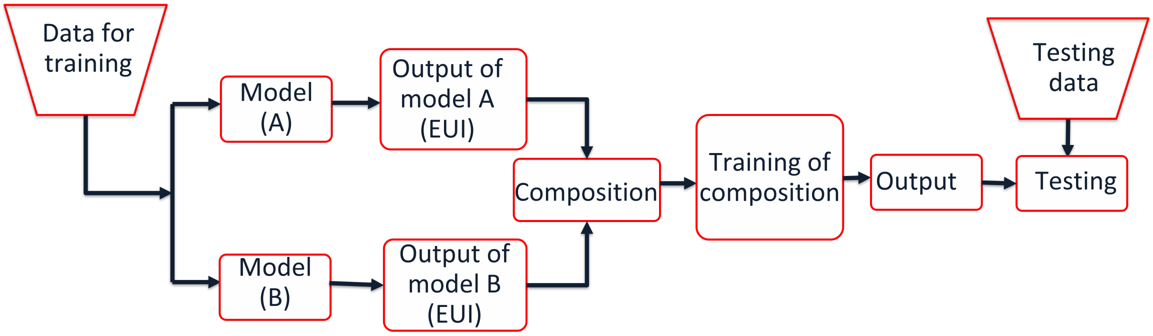


Figure 4.6: Composition procedure of component-based surrogate models.

4.6 Results

4.6.1 Sensitivity Analysis

A sensitivity analysis was done to predict the important building features from the geometry and envelope feature sets. For the sensitivity analysis, we used a common dataset which was simulated with 38,000 building geometry and 50 samples of 10 envelope features. The distribution of samples of features is shown in figure 4.7. The range of parameters is shown in figure 4.4 as well. Figure 4.8 shows a correlation heatmap between all 30 input features and the output (EUI), heating and cooling loads, the heating and cooling load intensities, and the total energy use (TEU). We simulated the models to get the EUI along with the heating and cooling load. The below helps to understand the relationship and impact of each feature with the output (EUI) along with heating and cooling loads. It shows that the most impactful geometry features in the geometry set for the total energy use are total surface area, perimeter, floor area, volume, the surface area of the bounding box, south and north wall areas, total wall area, and window areas. For EUI the most impactful was found to be the fenestration to floor area ratio. Similarly among the envelope features, the most impactful features were found to be roof insulation, solar transmittance (SHGC), wall insulation, and window to wall ratios for both EUI and TEU. For heating and cooling loads the most impactful features were perimeter, total surface area, total wall area, and total window areas. South, east, and west window areas, in particular, were found to be more correlated for cooling load. This information can be very helpful to consider suitable geometry and envelope features for fitting surrogate models to assist with energy-efficient building design.

4.6.2 Feature Selection

We defined a total of 30 geometry and building envelope features (20 geometry parameters and 10 envelope parameters). We then experimented to see if there were any unimportant features. First, we trained a geometry surrogate model with all 20 features, then decreased this to 16, 7, 5, 3, and 1 feature as shown in figure 4.9, and the performance of each feature set can be seen in figure 4.10. For the envelope features, we tried two combinations, one with a full set (10 features) and one without WWR (6 features) as shown in figure 4.11.

From this experiment, we found that for the geometry feature set the south, east,

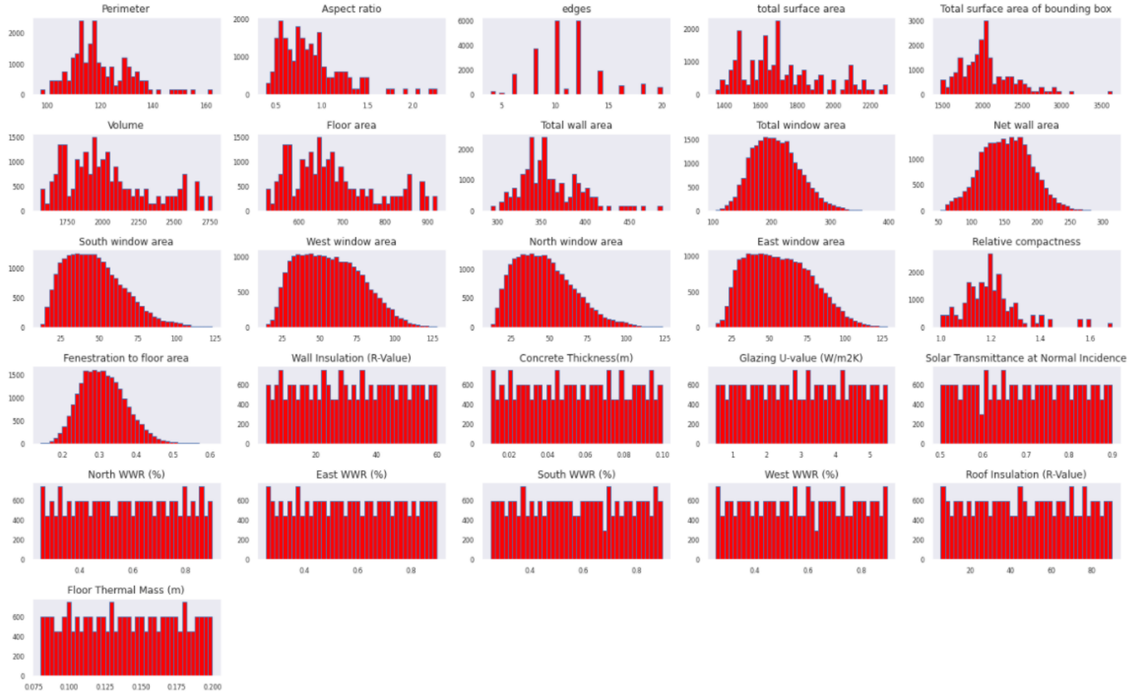
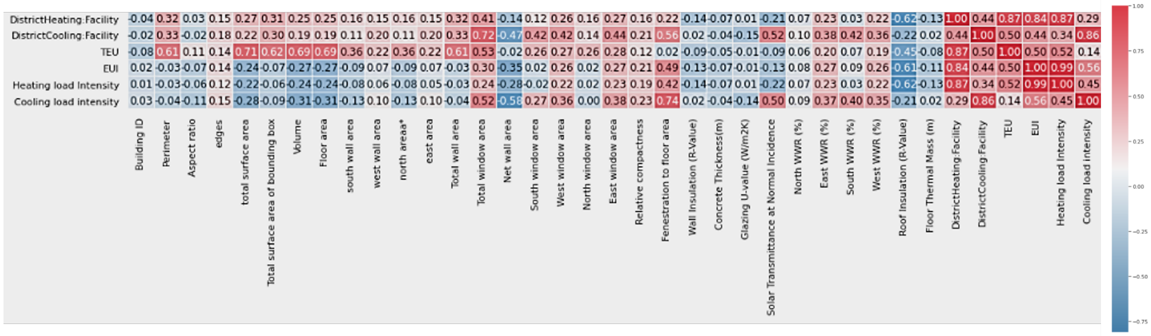


Figure 4.7: Distribution of feature samples for geometry and envelope parameters.



	Building ID	Perimeter	Aspect ratio	edges	total surface area	Total surface area of bounding box	Volume	Floor area	south wall area	west wall area	north areaa*	east area	Total wall area	Total window area	Net wall area	South window area	West window area	North window area	East window area	Relative compactness	Fenestration to floor area	Wall Insulation (R-Value)	Concrete Thickness(m)	Glazing U-value (W/m2K)	Solar Transmittance at Normal Incidence	North WWR (%)	East WWR (%)	South WWR (%)	West WWR (%)	Roof Insulation (R-Value)	Floor Thermal Mass (m)	DistrictHeating:Facility	DistrictCooling:Facility	TEU	EUI	Heating load Intensity	Cooling load intensity
DistrictHeating:Facility	-0.04	0.32	0.03	0.15	0.27	0.31	0.25	0.25	0.16	0.15	0.16	0.15	0.32	0.41	-0.14	0.12	0.26	0.16	0.27	0.16	0.22	-0.14	-0.07	0.01	-0.21	0.07	0.23	0.03	0.22	-0.62	-0.13	1.00	0.44	0.87	0.84	0.87	0.29
DistrictCooling:Facility	-0.02	0.33	-0.02	0.18	0.22	0.30	0.19	0.19	0.11	0.20	0.11	0.20	0.33	0.72	-0.47	0.42	0.42	0.14	0.44	0.21	0.56	0.02	-0.04	-0.15	0.52	0.10	0.38	0.42	0.36	-0.22	0.02	0.44	1.00	0.50	0.44	0.34	0.86
TEU	-0.08	0.61	0.11	0.14	0.71	0.62	0.69	0.69	0.36	0.22	0.36	0.22	0.61	0.53	-0.02	0.26	0.27	0.26	0.28	0.12	0.02	-0.09	-0.05	-0.01	-0.09	0.06	0.20	0.07	0.19	-0.45	-0.08	0.87	0.50	1.00	0.50	0.52	0.14
EUI	0.02	-0.03	-0.07	0.14	-0.24	-0.07	-0.27	-0.27	-0.09	0.07	-0.09	0.07	-0.03	0.30	-0.35	0.02	0.26	0.02	0.27	0.21	0.49	-0.13	-0.07	-0.01	-0.13	0.08	0.27	0.09	0.26	-0.61	-0.11	0.84	0.44	0.50	1.00	0.99	0.56
Heating load Intensity	0.01	-0.03	-0.06	0.12	-0.22	-0.06	-0.24	-0.24	-0.08	0.06	-0.08	0.05	-0.03	0.24	-0.28	-0.02	0.22	0.02	0.23	0.19	0.42	-0.14	-0.07	0.01	-0.22	0.07	0.23	0.03	0.22	-0.62	-0.13	0.87	0.34	0.52	0.99	1.00	0.45
Cooling load intensity	0.03	-0.04	-0.11	0.15	-0.28	-0.09	-0.31	-0.31	-0.13	0.10	-0.13	0.10	-0.04	0.52	-0.58	0.27	0.36	0.00	0.38	0.23	0.74	0.02	-0.04	-0.14	0.50	0.09	0.37	0.40	0.35	-0.21	0.02	0.29	0.86	0.14	0.56	0.45	1.00

Figure 4.8: Correlation heatmap showing the important features. The features with correlation factors shaded in blue have an inversely proportional relationship with the outputs and the features shaded in red are directly proportional to the output.

west, and north wall areas were not important as the R^2 score value of the model increased by 0.003 when we removed them. Hence, we used 16 features for further experiments. We found that WWRs were important for the envelope features as the model performance degraded from an R^2 score of 0.682 to -0.45 after we removed the four WWRs. Hence, we used all 10 features for further experiments.

4.6.3 Performance of different surrogate model formulations

The error metrics for each model formulation are shown in figure 4.14. Key points to note are:

- The performance of model A was the best among all the six models trained. It has an R^2 score of 0.993 MSE 0.006, and RMSE 0.0077 on common test data.

- The performance of model B was the second-best among all the six models trained. It has an R^2 score of 0.994, MSE 0.005, and RMSE 0.007 on common test data. The performance of model A and model B are almost identical because the information of the building given to both the models is the same but in a different format.

- For model C, the geometry sub-model had an R^2 score of 0.091 and the envelope sub-model had an R^2 score of 0.880. Finally, the final model after the composition of geometry and envelope sub-models had an R^2 score of 0.978, MSE 0.0203, and RMSE 0.142 on common test data.

- The geometry sub-model for model D had an R^2 score of 0.360 and the envelope sub-model had an R^2 score of 0.4300 and finally, the final model after the composition of two sub-models had an R^2 score of 0.386, MSE 0.708, and RMSE 0.842 on common test data. We see that the performance of Model D is the worst among all the other models trained. This might be because the common data set has variation in WWRs but the dataset we train our models (both geometry and envelope) does not capture this variation. This also shows that WWR is one of the main features influencing the Energy use intensity.

- The geometry sub-model for model E had an R^2 score of 0.982 and the envelope sub-model had an R^2 score of 0.973 and finally, the model after the composition of geometry and envelope sub-models had an score of 0.883, MSE 0.116, and RMSE 0.340 on common test data.

Geometry Features	10 Features	16 Features	7 Features	5 Features	3 Features	1 Features
Perimeter	✓	✓	✓			
Edges/Vertices	✓	✓	✓	✓	✓	
Total surface area	✓	✓	✓	✓	✓	✓
Total surface area of a bounding box	✓	✓				
Floor area	✓	✓	✓	✓	✓	
South wall area	✓					
North wall area	✓					
East wall area	✓					
West wall area	✓					
Window area (North)	✓	✓				
Window area (South	✓	✓				
Window area (East)	✓	✓				
Window area (West)	✓	✓				
Volume	✓	✓	✓	✓		
Total wall area	✓	✓				
Total window area	✓	✓				
Net wall area	✓	✓				
Relative compactness.	✓	✓	✓			
Fenestration to Floor Area Ratio (1 and 2)	✓	✓				
Aspect ratio	✓	✓	✓	✓		

Envelope Features	10 Features	16 Features
Wall Insulation (R-Value m^2·K/W)	✓	✓
Concrete Thickness(m)	✓	✓
Glazing U-value (W/m^{2}K)	✓	✓
Solar Transmittance at Normal Incidence	✓	✓
Roof Insulation (R-Value m^2·K/W)	✓	✓
Floor Thermal Mass (m)	✓	✓
North Window to Wall Ratio (%)	✓	
East Window to Wall Ratio (%)	✓	
South Window to Wall Ratio (%)	✓	
West Window to Wall Ratio (%)		

Legend	
🟥	Geometry parameter
🟦	Window parameter
🟩	Envelop parameter

Figure 4.9: Different combinations of feature sets for geometry and envelope sets.

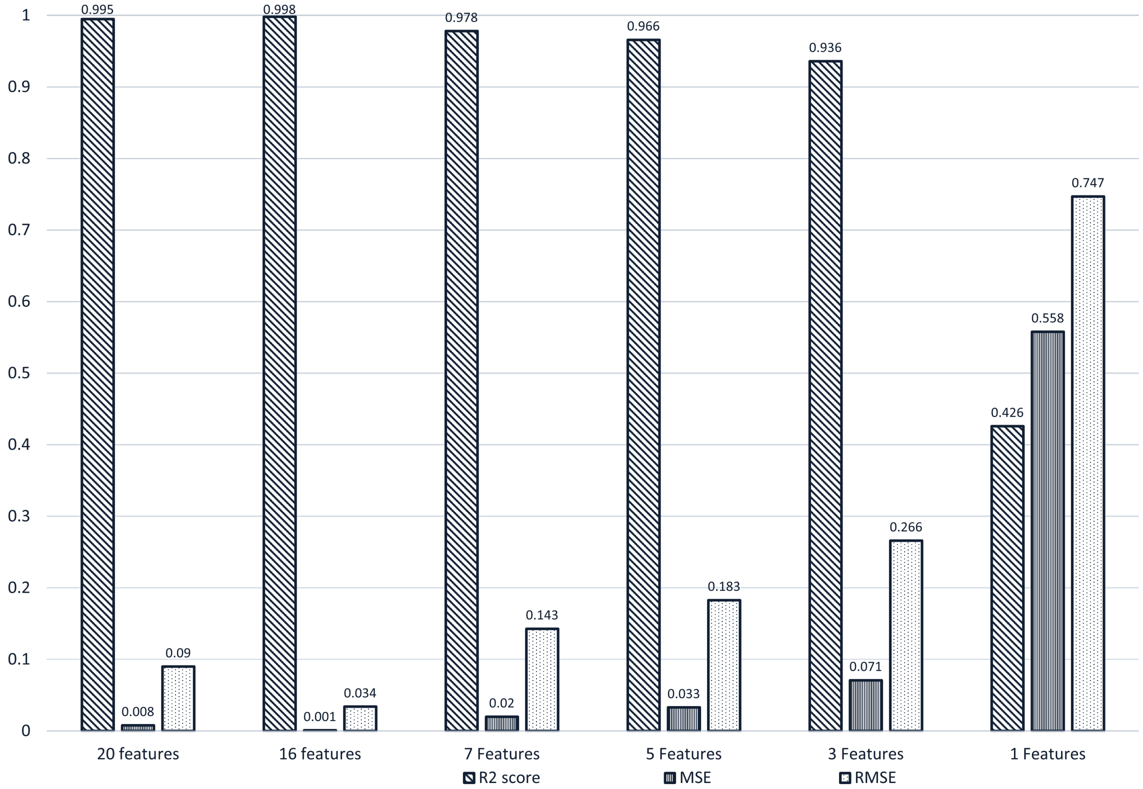


Figure 4.10: Performance of different geometry feature sets.

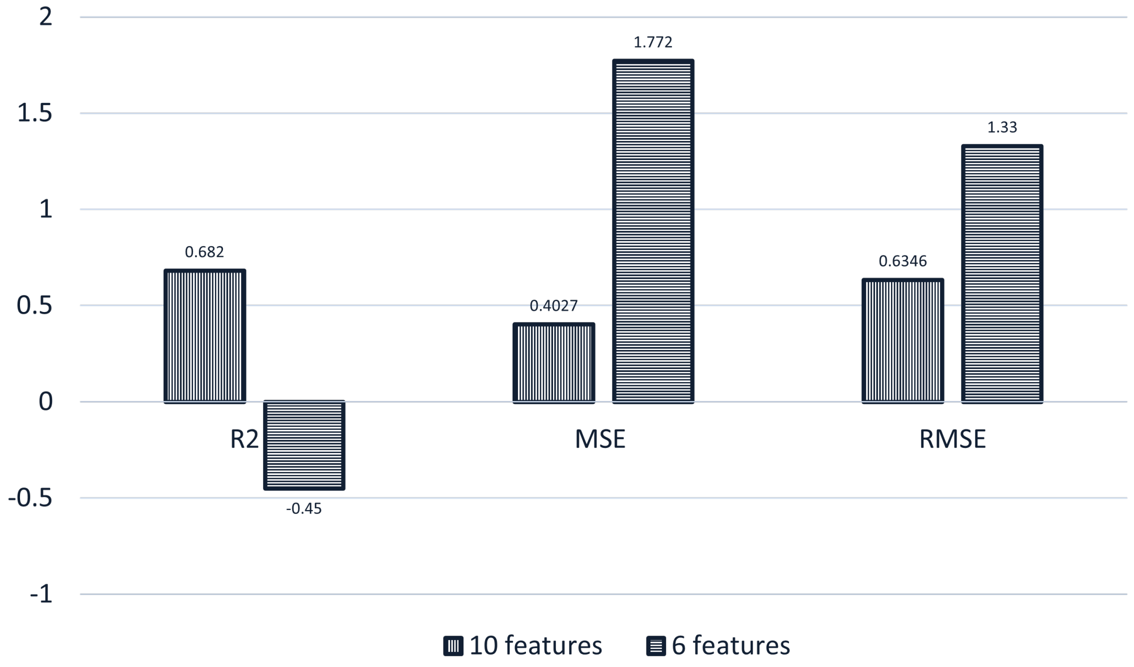


Figure 4.11: Performance of different envelope feature sets.

- For model F the geometry sub-model had an R^2 score of 0.988 and the envelope sub-model had an R^2 score of 0.953 and finally, the model after the composition of geometry and envelope had an R^2 score of 0.815, MSE 0.184, and RMSE 0.4300 on common test data.

There are a total of 6 final surrogate models trained in this research. The models were trained in different training sets but were tested on the same testing set. Among the six different models trained, we found that the monolithic models performed better than combined models, as would be expected. The best performing models among the 6 models were model A and model B, which had 19 and 22 inputs and one output. Both the models had similar information about the geometry and envelope of the building. Among the combined models, Model C performed the best. figure 4.12 shows the predicted and actual values of trained models, figure 4.13 is a plot made by overlapping the results from all the models to see the variance; we see that model D, model E, and Model F have more spread of values in comparison to other models. From the results, it is evident that a monolithic surrogate model with all the engineered features performs better than the one made by the composition of sub-models. But the performance of the composition models is also within an acceptable range. There are many benefits of a component-based surrogate model in comparison to a monolithic model with all the component features. The component-based surrogate model allows the surrogate model to be more versatile and convenient if we need to add more components to it, ultimately increasing the usability of the model. For example, if we need to add a feature set for shading, we can use a surrogate model trained on just shading features of the building and add it as another component rather than having to re-simulate the data including shading details and then training a monolithic model on it. On the other hand, if we use the monolithic model for different building components, we need to generate a new dataset and re-train the model each time we add a new component; this might also affect the importance of other features over the output.

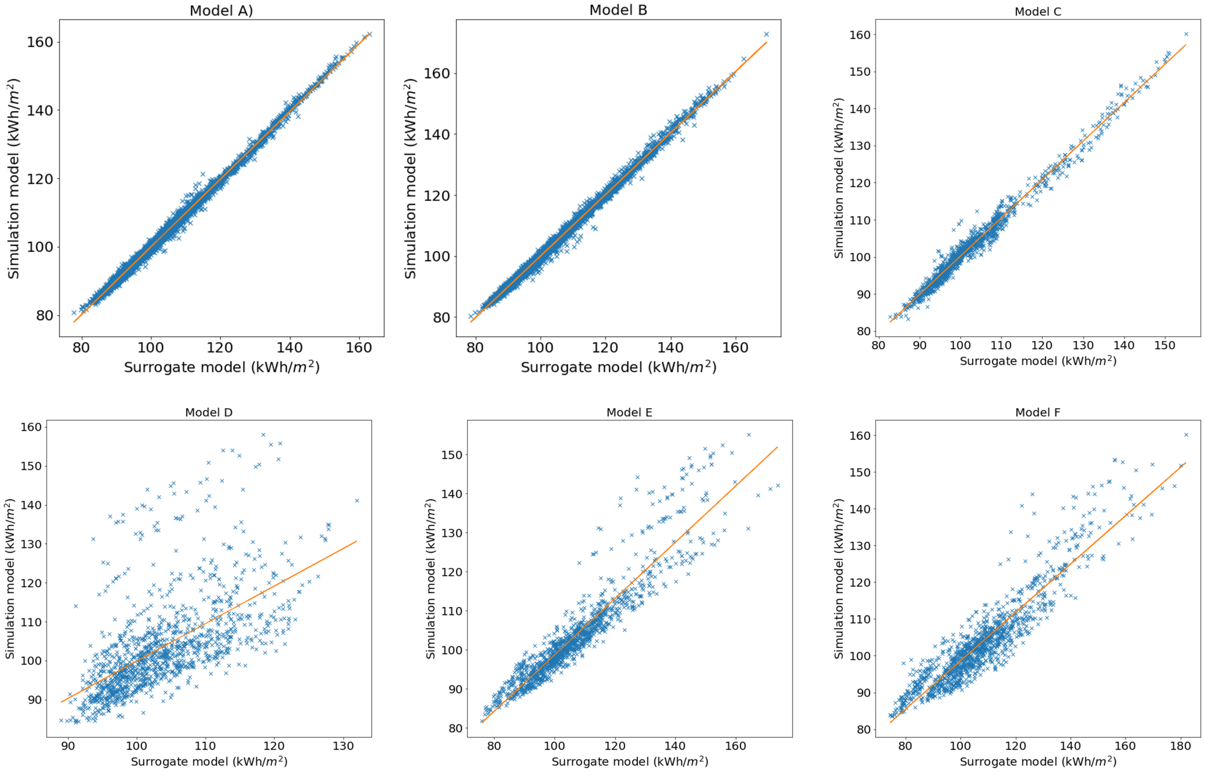


Figure 4.12: Performance of all the models.

4.7 Conclusion

This paper shows that machine learning-based surrogate models can very well capture the impact of geometry on the energy efficiency of buildings. The proposed off-the-shelf neural network-based surrogate model can be used as an effective tool for designing and analyzing energy-efficient designs at an early design phase as it saves a lot of computation time and resources. In this work, we develop six different surrogate models trained on various combinations of geometry and envelope features compared to performance in predicting energy use intensity (EUI). We used two types of surrogate modelling techniques, i.e., monolithic and component-based, and compared the performance. Three different performance metrics (R^2 score, MSE, and RMSE) were used to compare the accuracy of the models. The work confirms that expected monolithic models perform better, but the performance of the component-based model also falls under an acceptable range of accuracy for a conceptual design phase. In this stage, designers look for the impact of the features rather than the final

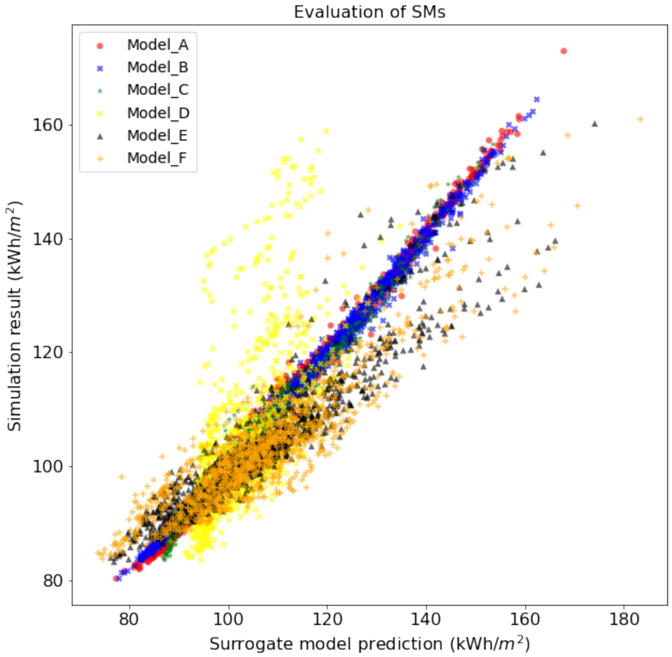


Figure 4.13: Performance of each model overlapped

energy output value hence the error from the model becomes almost negligible at this stage. Component-based surrogate models can be more efficient and flexible as having individual components, we can add or remove the set of features without having to regenerate and retrain the core model. We explored the statistical relationship of the 20 geometry features and 10 envelope features with the energy use of the building (EUI, TEU, heating load, and cooling load). We used a correlation heatmap for this purpose. The heatmap shows that individual wall areas were not that important but for EUI the most impactful geometry features were found to be the fenestration to floor area ratio, total window area, east and west WWR. Similarly, the most important geometry features for total energy use are perimeter, total surface area, the total surface area of the bounding box, volume, floor area, and total wall area. Similarly among the envelope features, the most impactful features were found to be roof insulation, solar transmittance (SHGC), wall insulation, and window to wall ratios for both EUI and TEU. The current setup forms the basis for further research where the complexities can be increased by adding components such as lighting, HVAC, etc. The workflow that extracts the 16 building features for geometry has

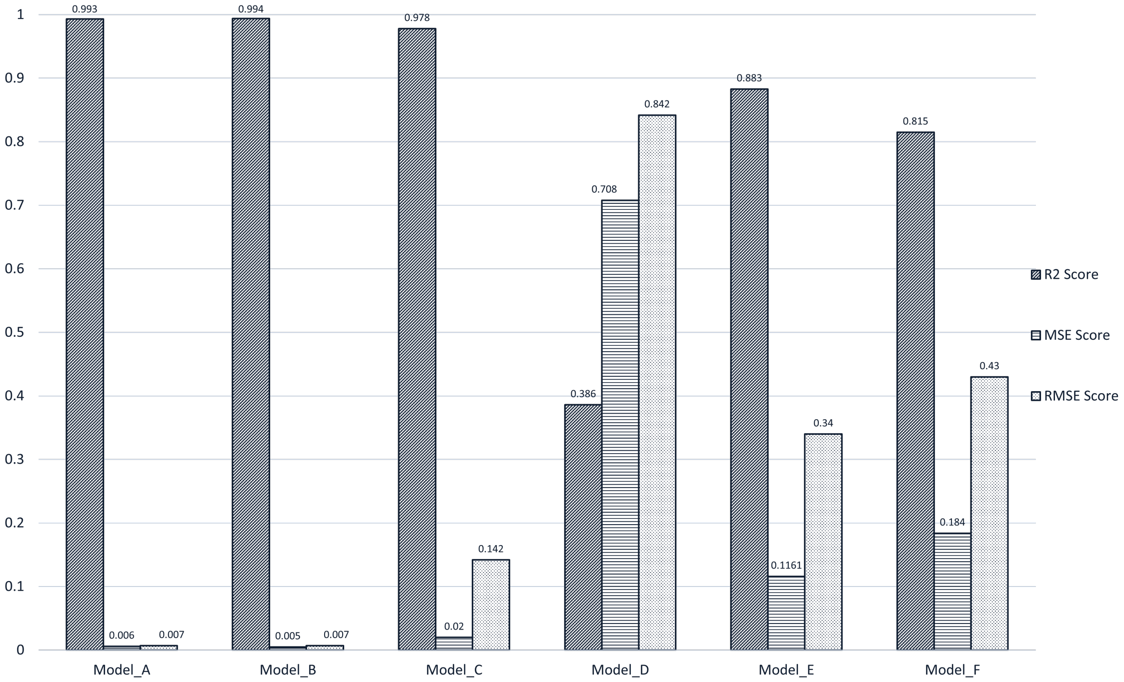


Figure 4.14: Error calculation for each surrogate model.

been developed in the GH environment. The workflow takes a 2D image of a floor plan and extracts the building features that are needed as an input for the surrogate model. The practicality of the model developed can be advanced by adding an image recognition feature that extracts the input features needed for the surrogate model from a basic floor plan or hand-sketched by the designer, allowing the designer to explore a wide range of designs space in no time. The ability of ML-based surrogate models to provide instant design feedback and information is crucial for the early stage of energy-efficient design. Surrogate models for sustainable building designs provide valuable information about the design parameters and their relationship with the output (in this case energy-use of the building), which will ultimately help improve the design from an early stage. The surrogate models developed in this research work can be used as a part of tools used in early design phases such as Net-zero navigator [67]. A platform such as Net-zero Navigator will help the building designers take advantage of surrogate modelling to drastically improve the performance of the building designs without learning the machine learning tools. The paper shows that there are a plethora of possibilities for exciting developments by combining machine-learning-based surrogate models with online visualization platforms. One of the other future works can be stimulating the data for different climate zones and teasing apart the heating and cooling load and seeing how the relationship with geometry and envelope changes according to the climate. More case studies and better design space sampling methods are required to evaluate the generalizability of these methods for their application in early design energy prediction.

Chapter 5

Conclusion and Future works

5.1 Conclusion

The goal of this thesis is to use surrogate models to analyze the impact of geometries on energy efficiency in buildings and attempt to identify important geometric features that affect the energy efficiency in buildings. Different surrogate models were trained to emulate the energy use of different types of building in different climates but primarily the work focused on residential buildings in Victoria, B.C. We have extracted up to 20 engineered geometric features that can very well represent the important shape features of the building along with 10 envelope features. For the variation in geometry of buildings, we developed a novel dataset of 38,000 building energy models and simulated them in various conditions by varying different parameters such as window to wall ratio, etc. and other envelope features such as U-values of the window, wall, roof, etc. and generated up to 5 different datasets to train multiple surrogate models to understand the impact of each feature with the energy output. We tried two different techniques, monolithic and component-based for training the surrogate models and compared their performance. We tried a component-based surrogate model due to its flexibility and reusability in the early design phase. We found that the monolithic model performs better than the component-based due to its simple model architecture but the component-based surrogate also performed reasonably well.

We did a sensitivity analysis for the inputs (geometric and envelope features) and output (total energy use, energy use intensity, heating, and cooling load). Based on the results of sensitivity analysis we found that the most important features for the total energy use were total surface area, floor area, volume, and the number of edges. We also found that the fenestration to floor area ratio was the most impactful feature while calculating the energy use intensity. Results show that with the increase in window area the energy use increases but we can achieve a low energy use with more windows if we stick to a lower number of edges in the building. From the results of the experiment done in this thesis, we can also conclude that with the increase in heating and cooling degree days the impact of geometry on the energy use of the building also increases.

Overall, we can conclude that a simple neural network-based surrogate model can be useful for the parametric optimization of geometry in buildings. Data-driven approaches are less time-consuming and less computationally expensive making them very useful to the designer at an early stage rather than using traditional building simulation tools. Using a geometry surrogate model for the conceptual phase of design gives the architect flexibility to explore various design options as they can get real-time energy information about the building without the burden of simulation of each design choice. Component-based surrogate models add an extra advantage to the use of machine learning models as they provide the flexibility to add various design components and make a whole building design analysis much easier and faster.

5.2 Future Works

The research work presented in this thesis has proposed the development of a quick energy prediction model using machine learning for the early stages. The current setup forms a basis for further research by adding more complexities such as making the models multizone, adding external shadings to the model, etc. Similarly, for the component-based surrogate modelling, we combined up to two sub-models, one for geometry and one for the envelope. The number of sub-models can be increased including HVAC, shading, etc which will make the model versatile and increase its usability for whole building design analysis. In near future, refined tools such as those presented in this paper can make a paradigm shift in the design process.

Appendix A

Appendix (Extension of Paper:2)

A.1 Impact of geometry across different climate zones

A.1.1 Introduction

Among various factors that building energy use depends on, climate/ location is one of the more prominent. The experiments done for paper 2 were based in Victoria, B.C., since the data used for sensitivity analysis and model training was simulated using the Victoria weather file. Here we extended this work to see how the impact of different geometric features changes in accordance with the change in climate data.

The results shown in the paper were also focused on the Victoria weather file so we extended the same work for different weather files across 6 different climate zones in Canada. The main objective of this extension was to analyze the impact of geometry on different weather data and see the impact of different weather on the relationship of different geometric features with the energy use intensity.

A.1.2 Methodology

For this experiment, we used the same workflow and the same set of varied geometry. We simulated the building energy models for 6 different cities each representing different climate zone from zone 4 - zone 8 as shown in A.1. The six different cities chosen were Victoria, Toronto, Ottawa, Edmonton, Whitehorse, and Yellowknife.

Data for heating and cooling degree days for all the six cities have been obtained from A.2. The workflow for generating the data is shown in A.3. We use the same

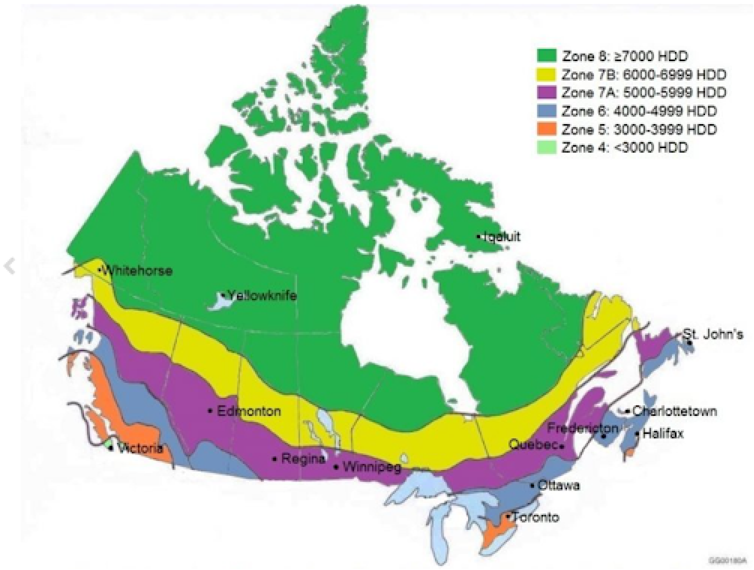


Figure A.1: Map of Canada showing 6 different climate zones.

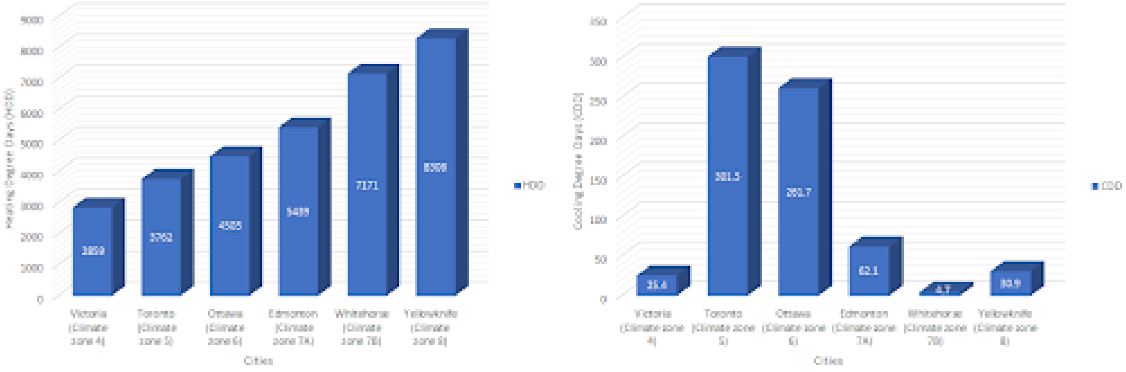


Figure A.2: Plot showing the number of heating degree days and cooling degree days for 6 different weather files.

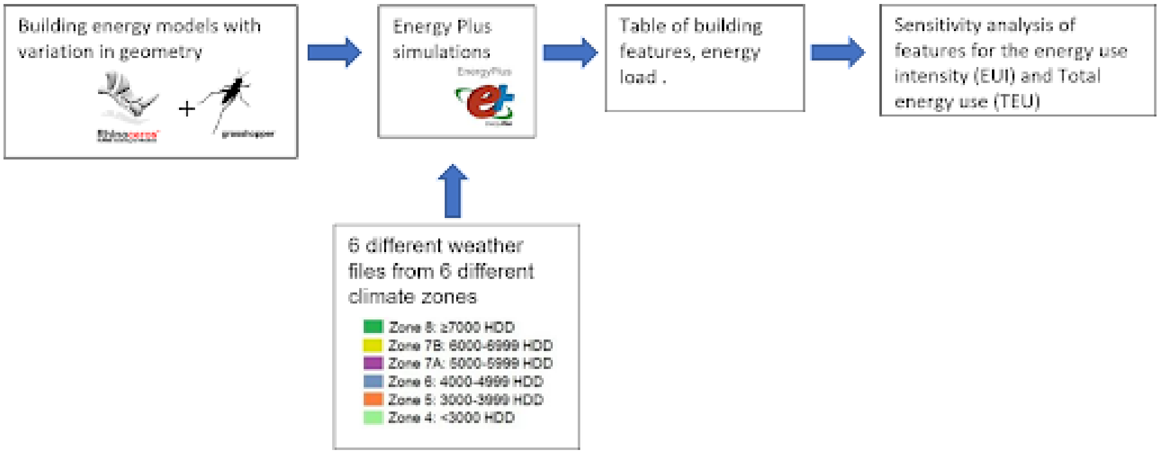


Figure A.3: Workflow for the experiment for 6 different climate zones across Canada.

set of building energy models and simulate them in six different weather files [4] from different climate zones. We then generate a table in an excel sheet with all the geometry features and energy use for each building and use that data for the sensitivity analysis geometry features and EUI.

A.1.3 Results

The figure A.4 shows the heating and cooling load of all the cities in a single plot. We see that heating load is highest in the cities with the most number of heating degree days and highest cooling load in the cities with the most number of cooling degree days. The highest cooling load was found to be in Toronto due to the most cooling load and extreme temperature in summer whereas Yellowknife had the highest heating load. Victoria weather was found to be the mildest weather in comparison to other cities. This shows that the experiment covers a wide variety of climatic conditions.

A.1.4 Sensitivity Analysis

The sensitivity analysis for the geometric features and energy use was done by plotting a correlation heatmap. From the analysis of the heatmap, it was evident that the most influential features for the energy use intensity were

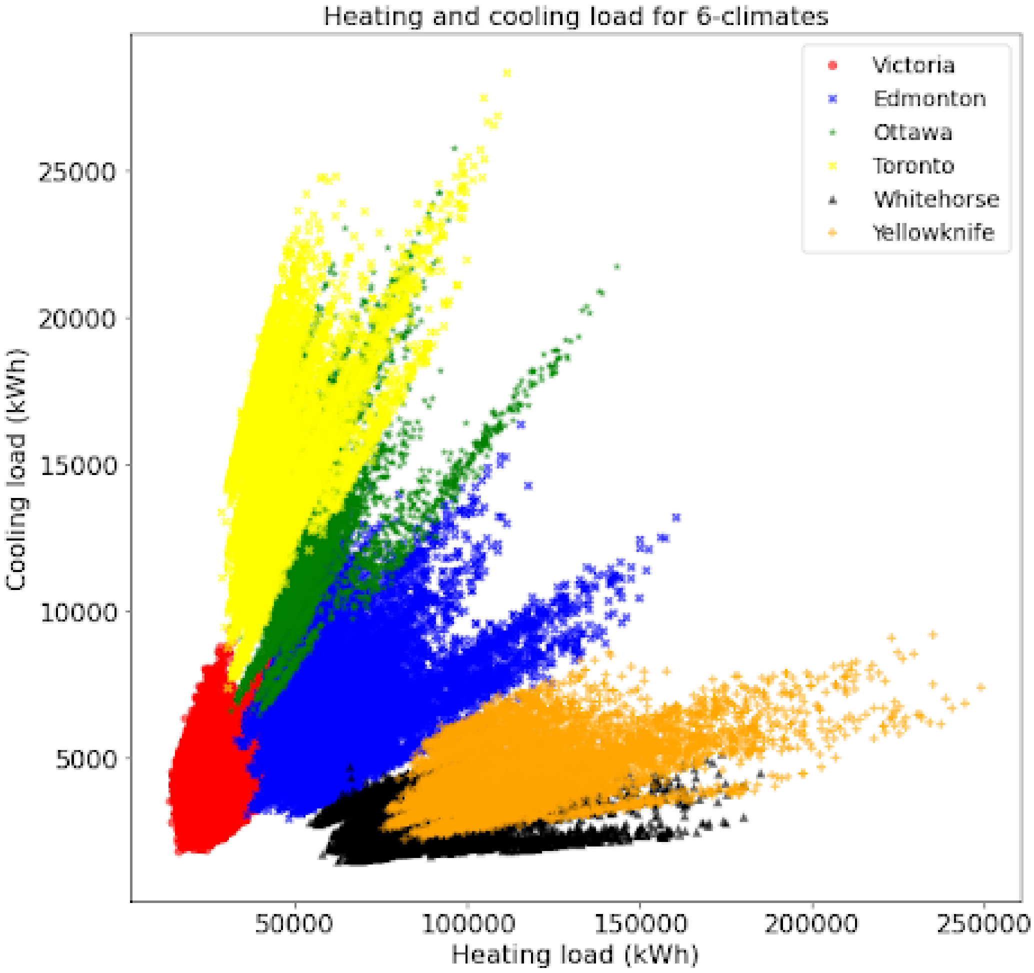


Figure A.4: Plot showing the heating and cooling load for 6-different cities.

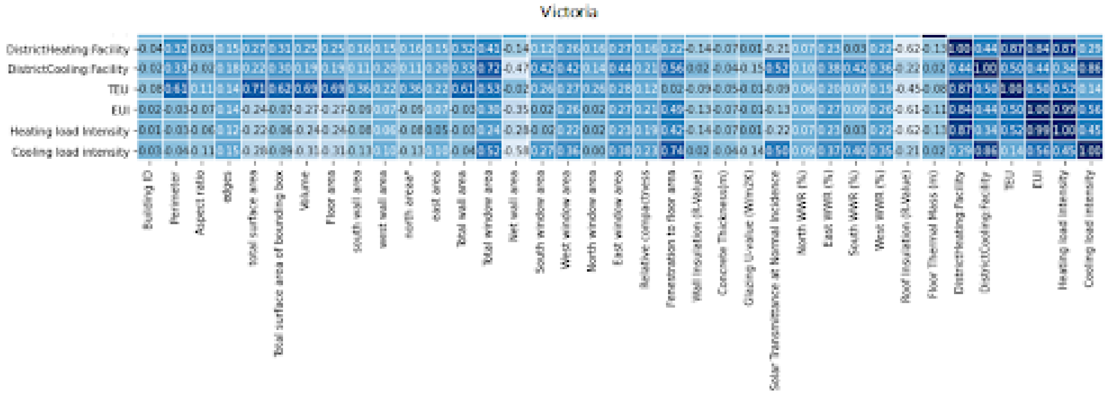


Figure A.5: Heatmap showing the Correlation of features with energy use for Victoria weather file

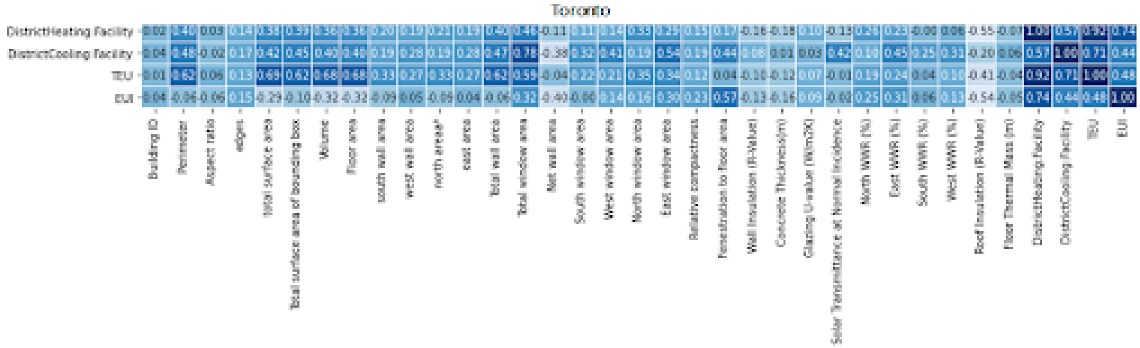


Figure A.6: Heatmap showing the Correlation of features with energy use for Toronto weather file

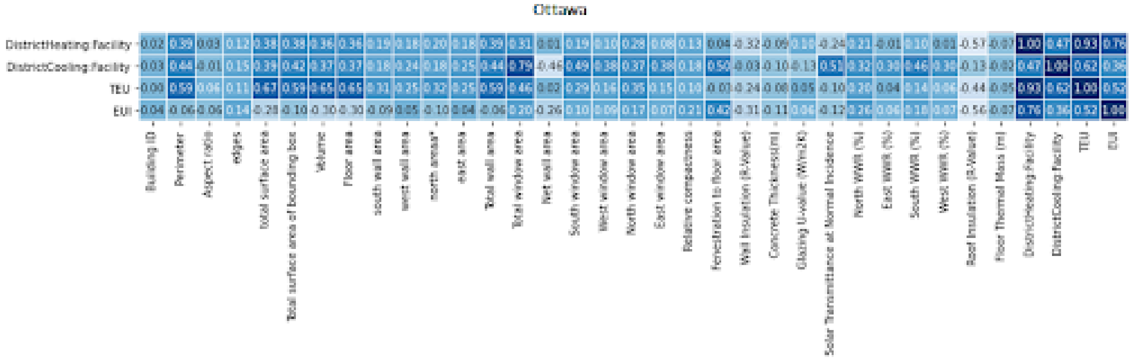


Figure A.7: Heatmap showing the Correlation of features with energy use for Ottawa weather file

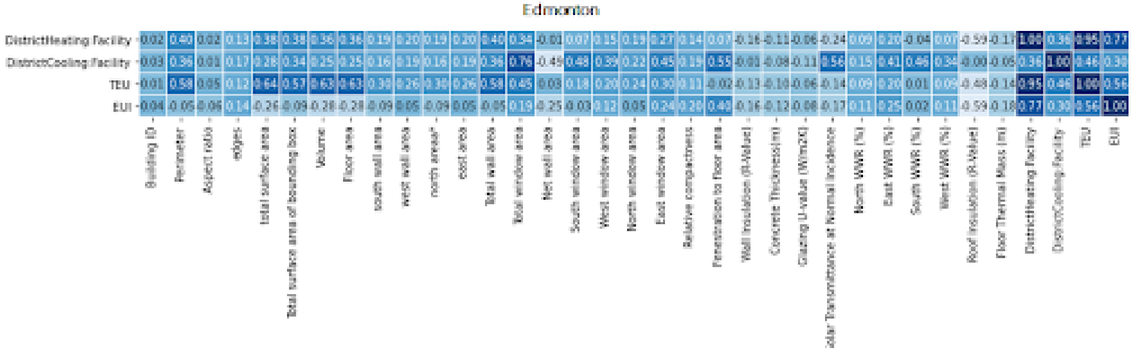


Figure A.8: Heatmap showing the Correlation of features with energy use for Edmonton weather file

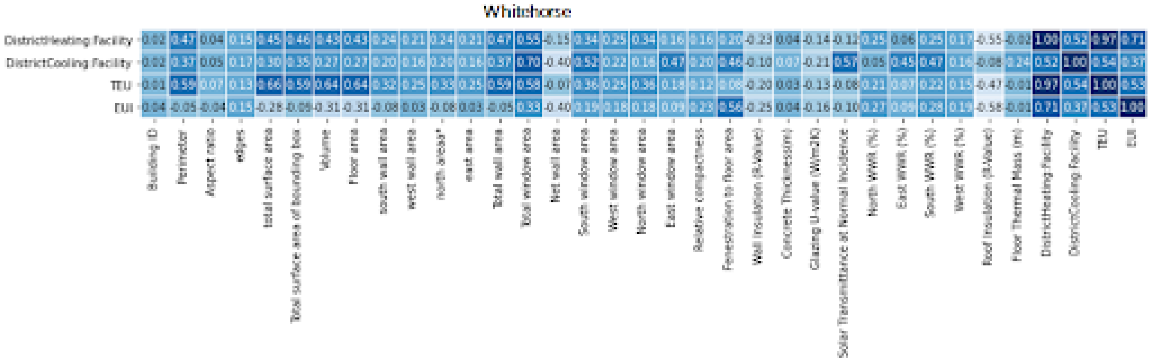


Figure A.9: Heatmap showing the Correlation of features with energy use for Whitehorse weather file

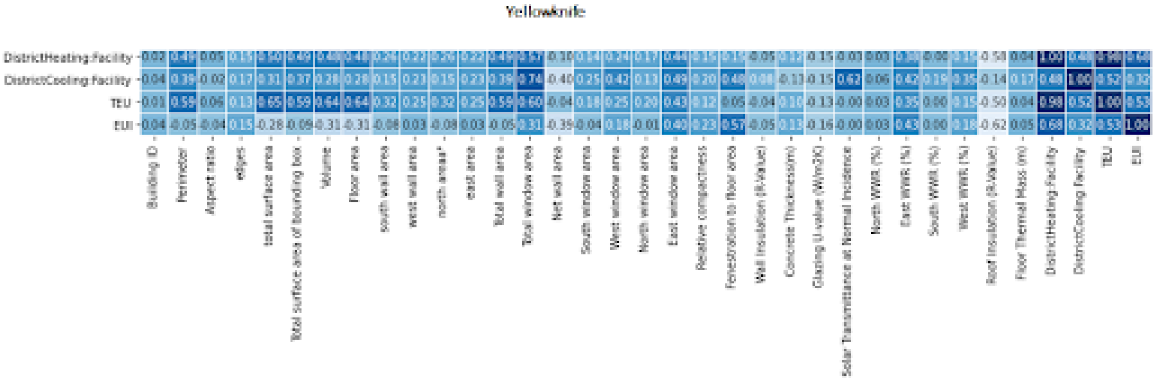


Figure A.10: Heatmap showing the Correlation of features with energy use for Yellowknife weather file

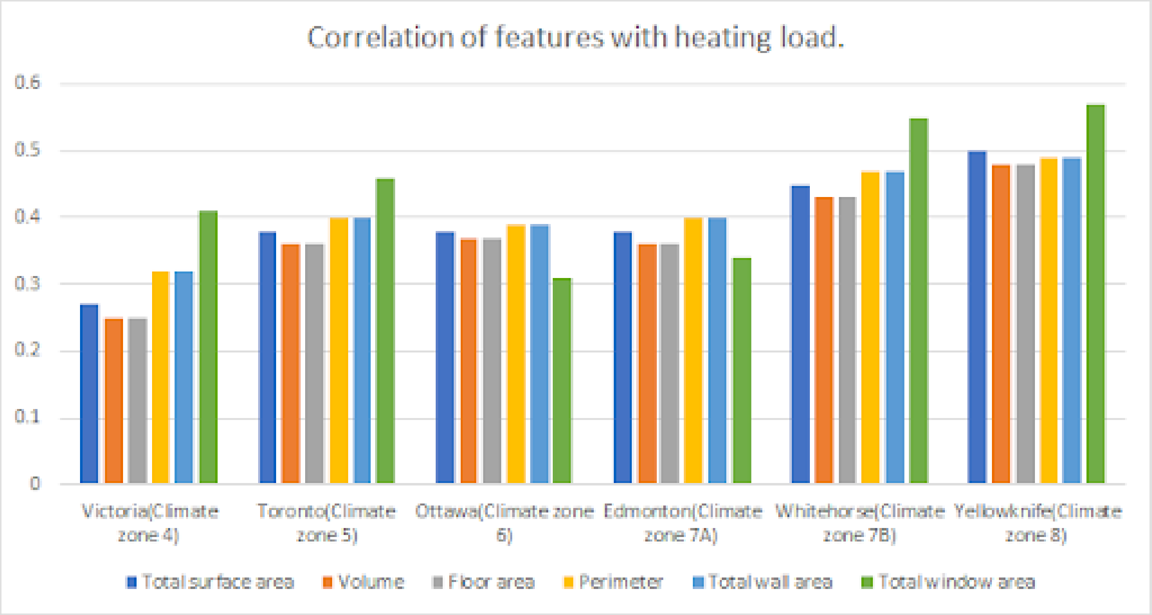


Figure A.11: Plot showing the correlation factor of features with the heating load.

A.1.5 Conclusion

From the analysis of the correlation between features and heating load A.11. We see that When the HDD increases the correlation factor for heating load, volume, floor area, perimeter, total wall area, and total window are increases. Meaning the impact of geometric features increases with an increase in the heating degree days for cold weather and Cooling degree days for hotter climates. From this experiment, we can conclude that the impact of geometry increases with the increase in heating and cooling degree days. This information can be helpful for the designers to design buildings in extreme weather conditions.

Bibliography

[1] MAPE-SCORE. https://www.statisticshowto.datasciencecentral.com,mean-absolute-percentage-error-mape/. Accessed: 2020-02-15.

[2] One-hot encoding. https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f. Accessed: 2020-02-15.

[3] R2-score. https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit. Accessed: 2020-02-15.

[4] Weather files. https://energyplus.net/weather. Accessed: 2020-02-15.

[5] Adnan AlAnzi, Donghyun Seo, and Moncef Krarti. Impact of building shape on thermal performance of office buildings in kuwait. *Energy Conversion and Management*, 50(3):822–828, 2009.

[6] Davide Anguita, Alessandro Ghio, Noemi Greco, Luca Oneto, and Sandro Ridella. Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In *The 2010 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.

[7] Somayeh Asadi, Shideh Shams Amiri, and Mohammad Mottahedi. 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, 2014.

[8] Shady Attia, Andre De Herde, Elisabeth Gratia, and Jan LM Hensen. Achieving informed decision-making for net zero energy buildings design using building performance simulation tools. In *Building Simulation*, volume 6, pages 3–21. Springer, 2013.

[9] Godfried Augenbroe. Trends in building simulation. *Building and Environment*, 37(8-9):891–902, 2002.

[10] John Basbagill, Forest Flager, Michel Lepech, and Martin Fischer. Application of life-cycle assessment to early stage building design for reduced embodied environmental impacts. *Building and Environment*, 60:81–92, 2013.

[11] Abdullatif E Ben-Nakhi and Mohamed A Mahmoud. Cooling load prediction for buildings using general regression neural networks. *Energy Conversion and Management*, 45(13-14):2127–2141, 2004.

[12] Bhumika Bhatta, Paul Westermann, and Ralph Evins. *Building simulation 2021 IBPSA*.

[13] Alexei Botchkarev. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. *Available at SSRN 3177507*, 2018.

[14] Yun-Tsui Chang and Shang-Hsien Hsieh. A review of building information modeling research for green building design through building performance analysis. *ITcon*, 25:1–40, 2020.

[15] Yixing Chen and Tianzhen Hong. Impacts of building geometry modeling methods on the simulation results of urban building energy models. *Applied Energy*, 215:717–735, 2018.

[16] Aidan Chopra. *Introduction to google sketchup.* John Wiley Sons, 2012.

[17] Jui-Sheng Chou and Dac-Khuong Bui. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings*, 82:437–446, 2014.

[18] Joseph Andrew Clarke and Jan LM Hensen. Integrated building performance simulation: Progress, prospects and requirements. *Building and Environment*, 91:294–306, 2015.

[19] Drury B Crawley, Linda K Lawrie, Curtis O Pedersen, and Frederick C Winkelmann. Energy plus: energy simulation program. *ASHRAE journal*, 42(4):49–56, 2000.

[20] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. Energyplus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4):319–331, 2001.

[21] Patrick Depecker, Christophe Menezo, Joseph Virgone, and Stephane Lepers. Design of buildings shape and energetic consumption. *Building and Environment*, 36(5):627–635, 2001.

[22] Bing Dong, Cheng Cao, and Siew Eang Lee. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37(5):545–553, 2005.

[23] Tomás Méndez Echenagucia, Alfonso Capozzoli, Ylenia Cascone, and Mario Sassone. The early design stage of a building envelope: Multi-objective search through heating, cooling and lighting energy performance analysis. *Applied energy*, 154:577–591, 2015.

[24] Betul Bektas Ekici and U Teoman Aksoy. Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, 40(5):356–362, 2009.

[25] Ralph Evins. A review of computational optimisation methods applied to sustainable building design. *Renewable and Sustainable Energy Reviews*, 22:230–245, 2013.

[26] Gaëlle Faure, Theo Christiaanse, Ralph Evins, and Gaby M Baasch. Besos: a collaborative building and energy simulation platform. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 350–351, 2019.

[27] Tonima Ferdous. Determining the effect of building geometry on energy use patterns of office developments. *Unpublished thesis*, 2012.

[28] Ivo ML Ferreira and Paulo JS Gil. Application and performance analysis of neural networks for decision support in conceptual design. *Expert Systems with Applications*, 39(9):7701–7708, 2012.

[29] A Fiksel, JW Thornton, SA Klein, and WA Beckman. Developments to the trnsys simulation program. 1995.

[30] Elisabeth Gratia and André De Herde. Design of low energy office buildings. *Energy and buildings*, 35(5):473–491, 2003.

[31] Rob Guglielmetti, Dan Macumber, and Nicholas Long. Openstudio: an open source integrated analysis platform. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2011.

[32] Debra Haltrecht, Radu Zmeureanu, and I Beausoleil-Morrison. Defining the methodology for the next-generation hot2000 simulator. In *Proceedings of Building Simulation'99*, volume 1, pages 61–68, 1999.

[33] Jung Min Han, Chih-Kang Chang, and Ali Malkawi. Arinet: Using 3d convolutional neural networks to estimate annual radiation intensities on building facades. 2020.

[34] Joshua Hester, Jeremy Gregory, and Randolph Kirchain. Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption. *Energy and Buildings*, 134:202–211, 2017.

[35] Zhijian Hou, Zhiwei Lian, Ye Yao, and Xinjian Yuan. Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique. *Applied energy*, 83(9):1033–1046, 2006.

[36] Janelle S Hygh, Joseph F DeCarolis, David B Hill, and S Ranji Ranjithan. Multivariate regression as an energy assessment tool in early building design. *Building and environment*, 57:165–175, 2012.

[37] IEA. Tracking buildings. Technical report, Internation Energy Agency, 2019.

[38] Soteris A Kalogirou and Milorad Bojic. Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy*, 25(5):479–491, 2000.

[39] Thulasi Ram Khamma and Mohamed Boubekri. Statistical analysis of impact of building morphology and orientation on its energy performance. *J. Eng*, 5:15–25, 2017.

[40] Bruno Lee, Navid Pourmousavian, and Jan LM Hensen. Full-factorial design space exploration approach for multi-criteria decision making of the design of industrial halls. *Energy and Buildings*, 117:352–361, 2016.

[41] Qiong Li, Qinglin Meng, Jiejin Cai, Hiroshi Yoshino, and Akashi Mochida. Applying support vector machine to predict hourly cooling load in the building. *Applied Energy*, 86(10):2249–2256, 2009.

[42] Mauricio Loyola. Big data in building design: a review. *ITcon*, 23:259–284, 2018.

[43] Ardeshir Mahdavi and Beran Gurtekin. Shapes, numbers, perception: Aspects and dimensions of the design-performance space. 2002.

[44] Wojciech Marks. Multicriteria optimisation of shape of energy-saving buildings. *Building and environment*, 32(4):331–339, 1997.

[45] Miguel Martin, Nyuk Hien Wong, Daniel Jun Chung Hii, and Marcel Ignatius. Comparison between simplified and detailed energyplus models coupled with an urban canopy model. *Energy and Buildings*, 157:116–125, 2017.

[46] RS Mitchell, JG Michalski, and TM Carbonell. *An artificial intelligence approach*. Springer, 2013.

[47] Thomas Olofsson, Staffan Andersson, and Ronny Östin. A method for predicting the annual building heating demand based on limited performance data. *Energy and Buildings*, 28(1):101–108, 1998.

[48] Torben Østergård, Rasmus L Jensen, and Steffen E Maagaard. Building simulations supporting decision making in early design–a review. *Renewable and Sustainable Energy Reviews*, 61:187–201, 2016.

[49] Josifas Parasonis, Andrius Keizikas, Audronė Endriukaitytė, and Diana Kalibatienė. Architectural solutions to increase the energy efficiency of buildings. *Journal of civil engineering and management*, 18(1):71–80, 2012.

[50] Werner Pessenlehner and Ardeshir Mahdavi. *Building morphology, transparence, and energy performance*. na, 2003.

[51] Adams Rackes, Ana Paula Melo, and Roberto Lamberts. Naturally comfortable and sustainable: Informed design guidance and performance labeling for passive commercial buildings in hot climates. *Applied Energy*, 174:256–274, 2016.

[52] Carlo Ratti, Dana Raydan, and Koen Steemers. Building form and environmental performance: archetypes, analysis and an arid climate. *Energy and buildings*, 35(1):49–59, 2003.

[53] Mostapha Sadeghipour Roudsari, Michelle Pak, Adrian Smith, et al. Ladybug: a parametric environmental plugin for grasshopper to help designers create an environmentally-conscious design. In *Proceedings of the 13th international IBPSA conference held in Lyon, France Aug*, pages 3128–3135, 2013.

[54] Stefano Schiavon, Kwang Ho Lee, Fred Bauman, and Tom Webster. Influence of raised floor on zone design cooling load in commercial buildings. *Energy and Buildings*, 42(8):1182–1191, 2010.

[55] Sundaravelpandian Singaravel and Philipp Geyer. Simplifying building energy performance models to support an integrated design workflow. In *International workshop of the european group for intelligent computing in engineering edition*, volume 23, 2016.

[56] Sundaravelpandian Singaravel, Philipp Geyer, and Johan Suykens. Component-based machine learning modelling approach for design stage building energy prediction: weather conditions and size. In *Proceedings of the 15th IBPSA conference*, pages 2617–2626, 2017.

[57] Sundaravelpandian Singaravel, Johan Suykens, and Philipp Geyer. Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction. *Advanced Engineering Informatics*, 38:81–90, 2018.

[58] Manav Mahan Singh and Philipp Geyer. *Building Simulation 2021 Ibosa*. IPBSA, 2021.

[59] Manav Mahan Singh, Sundaravelpandian Singaravel, Ralf Klein, and Philipp Geyer. Quick energy prediction and comparison of options at the early design stage. *Advanced Engineering Informatics*, 46:101185, 2020.

[60] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[61] PA Strachan, Georgios Kokogiannakis, and IA Macdonald. History and development of validation with the esp-r simulation program. *Building and Environment*, 43(4):601–609, 2008.

[62] Christian Struck, Pieter JCJ de Wilde, Christina J Hopfe, and Jan LM Hensen. An investigation of the option space in conceptual building design for advanced building simulation. *Advanced Engineering Informatics*, 23(4):386–395, 2009.

[63] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[64] Cihan Turhan, Tugce Kazanasmaz, Ilknur Erlalelitepe Uygun, Kenan Evren Ekmen, and Gulden Gokcen Akkurt. Comparative study of a building energy performance software (kep-iyte-ess) and ann-based building heat load estimation. *Energy and Buildings*, 85:115–125, 2014.

[65] Weimin Wang, Radu Zmeureanu, and Hugues Rivard. Applying multi-objective genetic algorithms in green building design optimization. *Building and environment*, 40(11):1512–1525, 2005.

[66] Paul Westermann and Ralph Evins. Surrogate modelling for sustainable building design-a review. *Energy and Buildings*, 2019.

[67] Paul Westermann, David Rulff, Kevin Cant, Gaelle Faure, and Ralph Evins. Net-zero navigator: A platform for interactive net-zero building design using surrogate modelling paul westermann1, david rulff1, kevin cant1, gaelle faure1, ralph evins1 1energy in cities group, department of civil engineering, university of victoria, canada.

[68] Michael Wetter. Modelica-based modelling and simulation to support research and development in building energy and control systems. *Journal of Building Performance Simulation*, 2(2):143–161, 2009.

[69] Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhan Wang, Yu-Hao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 38(6), 2019.

[70] Chunhai Xia, Yingxin Zhu, and Borong Lin. Building simulation as assistance in the conceptual design. In *Building simulation*, volume 1, pages 46–52. Springer, 2008.

[71] Zhun Yu, Fariborz Haghighat, Benjamin CM Fung, and Hiroshi Yoshino. A decision tree method for building energy demand modeling. *Energy and Buildings*, 42(10):1637–1646, 2010.

[72] Izzet Yüksek and Tülay Tikansak Karadayi. Energy-efficient building design in the context of building life cycle. *Energy Efficient Buildings*, page 93, 2017.

[73] Gabriela Zapata-Lancaster and Christopher Tweed. Tools for low-energy building design: an exploratory study of the design process in action. *Architectural Engineering and Design Management*, 12(4):279–295, 2016.