Surrogate Modelling of Solar Radiation Potential for the Design of PV Module Layout on Entire Façade of Tall Buildings

Meggie Vincentia Barus

Master's Thesis

Department of Construction Management and Engineering

University of Twente, 2021

Abstract:

This research investigated the performance of a surrogate modelling approach for the simulation of solar radiation potential on the vertical surfaces of tall buildings. Surrogate modelling was used to approximate the input-output behaviour of the existing simulation model. The Random Forest (RF) machine learning approach was used to investigate three different scenarios, namely (1) Random variation, (2) Grid variation, and (3) Uniform variation, and a Genetic Algorithm was used as the hyperparameter optimisation. A case study using a building in Sir George William (SGW) campus of Concordia University in downtown Montreal Canada was performed to investigate the performance of surrogate models. As a result, even by only using a small sample size of the dataset when developing the RF, surrogate modelling can give 94% accuracy to approximate the simulation of solar radiation. From the three scenarios, the best accuracy is obtained when using the Random variation method. In short, the solar radiation simulation is very complex and too sensitive to the location and shadow effect. Therefore, simplification of those factors cannot be made to approximate the solar radiation potential. Also, using RF, the computational time improved by 16 times faster than when using the existing simulation model.

Keywords: Surrogate modelling, machine learning, genetic algorithm, solar radiation, vertical surface

1. Introduction

The tall building sector has approximately 32% share of total energy demand throughout the world (Khatib, 2012). These buildings are responsible for the annual consumption of 40% of the total energy used across the developed countries (Yüksek & Karadayi, 2017). Not to mention, a large portion of these energy still uses fossil-based energy source, which is inevitably limited and depleting. However, the worldwide energy consumption records show that the electrical energy demand will continuously increase (IEA, 2018). This rise occurred due to the forecasted growth of the world population, leading to a vast number of new buildings. In the long term, it would not be possible to meet this consumption demand by supplying energy to various locations from a centralised source of energy. Therefore, in recent years, researchers and building planners have begun to focus on creating decentralised energy generation where each building can (partly) supply its own energy (Marszal, et al., 2011).

Photovoltaic (PV) solar energy is one of the best sources of clean energy used in the building environment to substitute fossil-based energy partially and can be used to supply local energy demand (Kåberger, 2018). This energy is harvested by installing solar panels on the exterior of buildings. Solar panels, also known as PV panels, are commonly installed on the building's rooftop or any horizontal or tilted surfaces. Nevertheless, the current market of PV panels allows the installation on various surfaces of building surfaces. For instance, building-integrated photovoltaics (BIPV) enables solar energy harvesting from buildings' façade, as shown in Figure 1. BIPVs can reduce the overall material cost because they serve multiple functionalities (Raugei & Frankl, 2009; Jelle et al., 2012). This means vertical surfaces, too, have the opportunity to produce a high amount of energy if panels are strategically installed on them (Catita et al., 2014). It is shown that PV panel installation on the vertical surfaces of tall buildings is promising (Liang et al., 2014; Salimzadeh, et al., 2020). The effective use of vertical surfaces for harvesting clean energy on a tall building is essential. After all, it is shown

that the average height of buildings in an area has a negative impact on energy consumption because it increases the population density (Resch, et al., 2016; Godoy-Shimizu, et al., 2018).

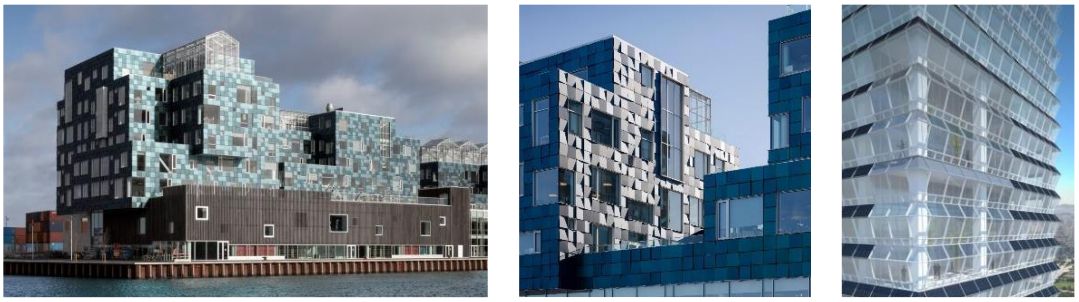


Figure 1 Examples of using PV modules on Facades (Gibson, 2017; Smith & Gill, 2014)

Despite the potentials, vertical surfaces of tall buildings are seldom leveraged for harvesting solar energy. A study conducted over an urban area in Madrid showed that despite the vast surface areas that can be used for PV installation, only 7.22% of the area's façade is considered usable for the photovoltaic system. This number is based on the current practices and regulations on the minimum thresholds required for PV installation (Esclapés, et al., 2014). Numan et al. (2020) identified the barriers against the maximum use of PV panels. They concluded that one obstacle is that tall buildings are usually located in dense residential or commercial areas, where many other buildings surround them. Because of the shadow effect of buildings on one another, vertical surfaces receive considerably less solar radiation than horizontal surfaces. That is why the efficiency of vertical solar panels is heavily dependent on how their layout is designed (Numan, et al., 2020). In other words, the economically viable use of vertical surfaces of tall buildings requires strategic placing and spacing of panels on the vertical surfaces, i.e., the layout design (Salimzadeh, et al., 2020).

To assess the potentials of different vertical surfaces for radiation harvesting and to find the most efficient PV's layout design, a detailed solar simulation of the building surfaces considering the surroundings is required (Gurupira & Rix, 2017). The PV layout is typically determined by multiple factors such as the location, sizes, and orientation of panels (Middelhauve, et al., 2021).

There are many different approaches for the simulation and assessment of solar radiation potentials of urban surfaces (Hwang, et al., 2012; Kucuksari, et al., 2014; Freitas, et al., 2015; Koo, Hong, Lee, & Kim, 2016). Conventionally, physics-based numerical models, e.g., Atmospheric and Topographic Model (ATM), were used to assess the solar potential of surfaces in an urban environment (Paulescu, et al., 2012). But these earlier methods were only applicable on large scales and could not support detailed surface-level analysis. Later, and with the rising popularity of Geographical Information Systems (GIS), solar radiation analysis could be done on a more granular level using 2D models developed based on Digital Elevation Model (DEM), Light Detection and Ranging (LiDAR), and photogrammetric approaches (Kumar, et al., 1997; Jochem, et al., 2009; Chow, Fung, & Li, 2014). However, GIS-based methods of solar radiation analysis were limited to horizontal surfaces or flat rooftops because they were only using 2D and 2.5D models, which lacked the sufficient level of 3D detail (Carneiro, et al., 2010; Esclapés, et al., 2014).

This situation has changed in recent years with the advent of 3D modelling methods (Brown, 2016). Nowadays, software packages like RADIANCE can perform solar analysis on volumetric 3D models and curved geometries using accurate ray-tracing algorithms (Ward, 1994). In fact, more sophisticated methods and tools such as Cumulative Sky and Daysim can perform by considering climate data, shading factors, and surrounding buildings (Mardaljevic, 2000; Robinson & Stone, 2004). Although these tools and methods use 3D models as an input to consider the shading effect, they still take an indiscriminate approach towards different surfaces. They, therefore, are not amenable for the analysis of PV modules that can only be used on specific types of surfaces. For instance, BIPV modules or transparent PV modules that can be used on windows. This limitation is a significant factor, especially for installing PV modules on vertical surfaces, because the diversity of vertical surfaces (e.g., windows, walls, curtain walls, balconies) requires surface-specific simulation of surfaces.

With the improvement of PV modules technology and supporting simulation tools, researchers started to consider the vertical surfaces of the buildings in the urban area for solar radiation analysis (Gooding, et al., 2013; Esclapés, et al., 2014; Martín, et al., 2015). In recent years and with the rising popularity and availability of Building Information Models (BIM), it has become possible to leverage the semantic data embedded in the 3D model of buildings to perform surface-specific simulation of building surfaces. Several recent studies have demonstrated this possibility (Ning et al., 2018; Al-Janahi et al., 2020; Salimzadeh, et al., 2020). For instance, the authors have developed a parametric model that can generate a wide range of PV layout alternatives considering the surface type and layout of the module (Salimzadeh, et al., 2020). In this model, the user can specify the value of pre-defined parameters (e.g., number and location of panels, the tilt angles, etc.) to generate and assess a specific design alternative. However, on their own, these parametric models are not sufficient to identify the best layout design. It is not enough because the design space of this optimisation problem can become so large that an exhaustive search of all possible alternatives becomes practically and computationally very challenging. Therefore, it is imperative to integrate these parametric models with an optimisation method to find the (near) optimum design using a more efficient search strategy. This framework of integrating a parametric model with an optimisation strategy for design problems is commonly referred to as simulation-based optimisation or generative design (Nguyen et al., 2014).

Given the extended size of the design space in these simulation-based optimisation problems, meta-heuristic optimisation methods, e.g., Genetic Algorithm (GA), are often considered (Wang & Schafer, 2020; Thrampoulidis et al., 2021). Many randomly selected initial solutions are evaluated and transformed using evolutionary operations in these optimisation methods to find a near-optimal solution. While proven to be effective, the metaheuristic models are sensitive to several parameters. Most importantly, the optimality of the final result depends

heavily on the evaluated number of solutions, i.e., the higher the number of solutions, the greater the chance of optimality. The authors have previously presented a generative design framework for PV layout design by integrating a BIM-based parametric model and GA (Salimzadeh N. , 2021).

Nevertheless, the main problem with the current generative design framework is that the parametric modelling platform is inherently physics-based. In other words, for each design alternative, the simulation needs to perform the ray-tracing approach for every single panel and for every day to assess the amount of annual solar radiation potential. This makes the simulation platform computationally expensive. When combined with a meta-heuristic optimisation method, where several generations of a large population of design alternatives need to be explored, the parametric platform can be inefficient because of the computation time. This is especially important because PV layout optimisation is only one of many criteria that need to be considered for the optimal design of building facades. Other criteria, such as insolation and aesthetics, also need to be considered. Therefore, in the general practice of building design, having a computationally very expensive pipeline for the PV layout design may push designers to forgo the use of the analytical approach for the façade design and restore to heuristic-based methods.

One potential approach to address this problem is to use Machine Learning (ML) models that can substitute the computationally expensive simulation model. This substitute model is known as a meta-model or surrogate model (Arisha & Abo-Hamad, 2010). As shown in Figure 2, a machine-learning model can learn from a large volume of training data and identify relationships and patterns between a set of dependent and independent variables, i.e., inputs and outputs of the simulation model (Elfaki et al., 2014). Surrogate models are shown to be very useful for reducing the computational intensity of the simulation-based optimisation problem (Karnon, et al., 2012), because they can reduce the computational intensity of the

optimisation through an accurate mathematical approximation of the physics-based simulation models. Bornatico et al. (2013) applied surrogate modelling to optimise PV systems and demonstrated that the computation time could be reduced to 150 times less than physics-based simulation. Similar results were reported by Perera et al. (2019). However, this study only considered the optimisation of the PV system from the mechanical perspective and not the layout design. Xu et al. (2017) also showed promising results in applying surrogate models for the building façade design. However, only the impact of using a default BIPV system on the cost of the overall design was considered and not the detailed layout design.

To the best knowledge of the authors, the surrogate modelling approach on design optimisation of solar panel layout on the building façades has not been considered before.

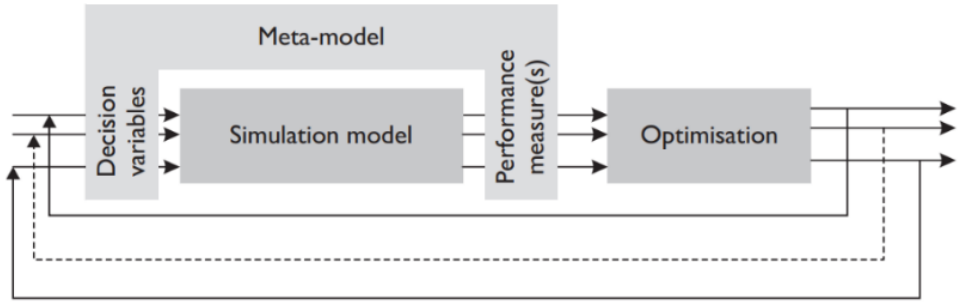


Figure 2 Schematic representation of meta-modelling

1.1. Research Objective and Scope

On the premise of the above research gap, this study aims to investigate the feasibility and effectiveness, i.e., in terms of prediction accuracy, of using a meta-model to assess the solar radiation potential of building façade. To this end, first, a framework will be developed to apply the concept of surrogate modelling for the approximation of the PV layout parametric model developed previously by Salimzadeh et al. (2020). Different approaches for the development of these surrogate models will be explored. Finally, the performances of the other surrogate models are assessed through comparison with the results of the parametric model.

This newly developed framework is expected to provide the building façade designers with an insight into how data-driven metamodeling techniques can help incorporate a better analytical approach in building design. It should be highlighted that this research focuses only on developing the surrogate model for the vertical surfaces of the building, and thus the rooftop PV modules are not considered. Nevertheless, it is expected that the same method can be applied to the rooftop as well.

The remainder of this paper is organised as follows. Section 2 presents the proposed framework and various metamodeling approaches. Section 3 elaborates on implementing the proposed method into a case study and is followed in Section 4 by the corresponding results. Finally, the conclusions, limitations, and future work are presented in section 5.

2. Proposed method

This chapter presents the overall method applied in this research to develop the surrogate model of solar radiation potential. Figure 3 shows the overview of this method. Overall, the proposed method consists of three phases. The first phase is allocated to building the dataset that will be used to develop the machine learning model. In this phase, several datasets will be generated based on different strategies for approximating solar panel behaviour. Next, in Phase 2, an ML method is used to develop the surrogate model for other datasets. Then, the performance of the developed surrogate model is evaluated through comparison with the physics-based simulation model.

2.1. Building Datasets

As shown in Figure 3, the first phase is dedicated to building the dataset used for training and validation of the surrogate model. To this end, a previously developed parametric model of the façade PV module is used (Salimzadeh, et al., 2020). For completeness, a brief overview of this parametric model is explained below.

2.1.1. Simulation-based Parametric Model

Figure 4 shows a schematic representation of how the PV module layout parametric model functions. Figure 4(a) shows that the potential candidate location for installing PV modules is first selected on the external surfaces. In this step, only feasible surfaces are considered. Therefore, if there are elements on a specific surface, e.g., a mechanical unit, that hampers the installation of the PV module, the surface is excluded. Next, the parametric model allows the user to specify the geometric specification of the PV module in terms of width (W_i), length (L_i), and tilt angle (θ_i), as shown in Figure 4(b). Finally, the user-specified layout is determined in terms of panels that are installed on the desired location (P_i), as shown in Figure 4(c).

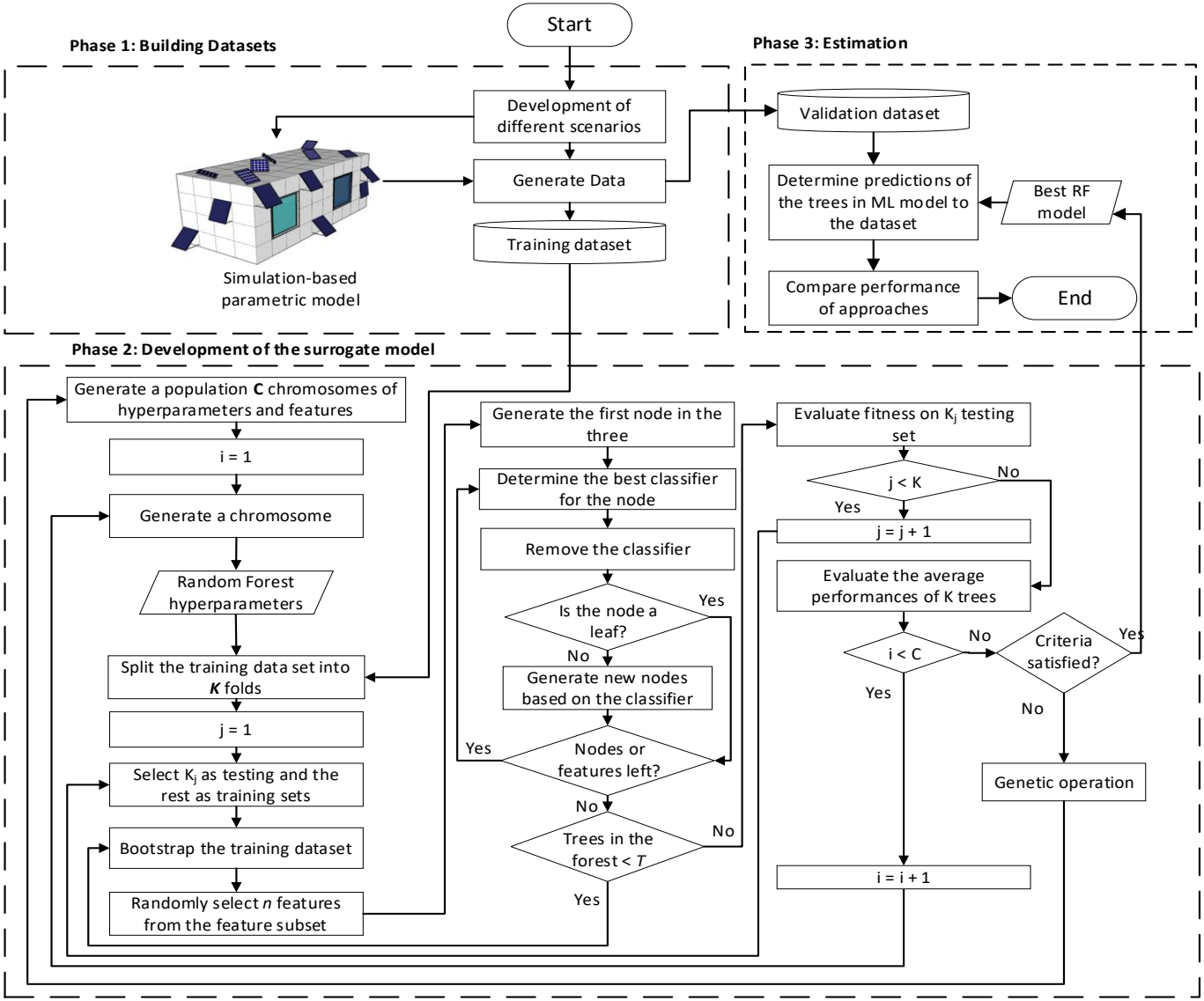


Figure 3 The framework of the research method

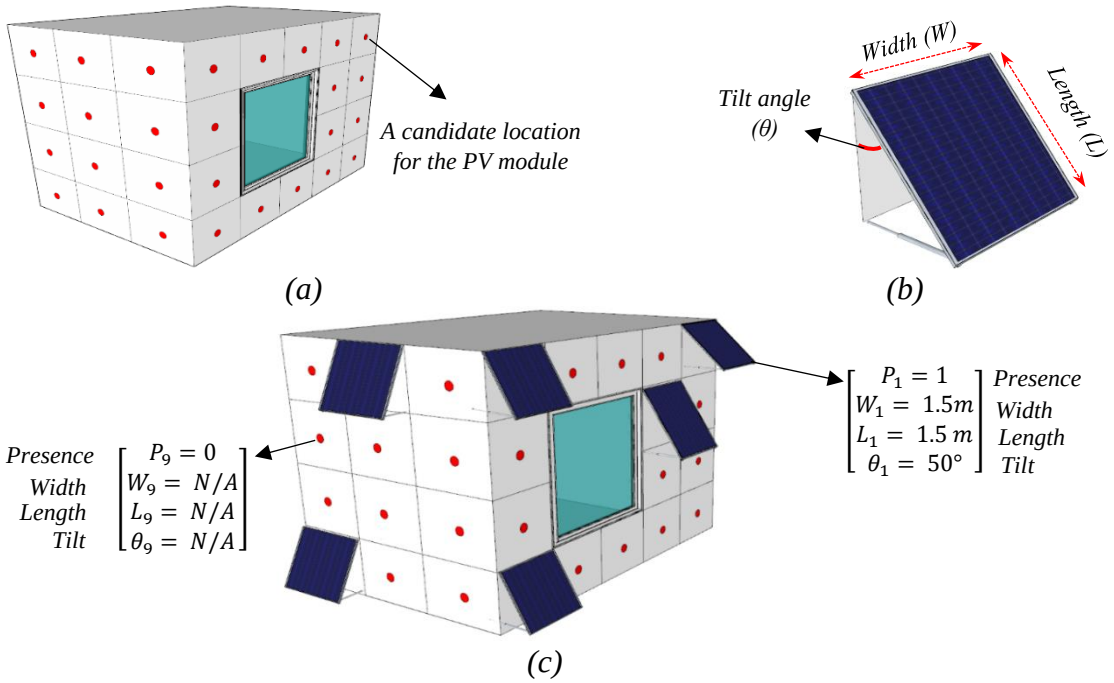


Figure 4 Simulation-based parametric model of PV module layout

2.1.2. Data Structure for Machine Learning

In preparing the data for ML development, it is vital to identify the relevant features that will be used for the training of the ML model. The selection of features determines the parameters deemed to be suitable for predicting the PV module radiation potential. In this research, the following features are considered: (1) *location of the PV Module*: the location of the PV module, as shown in Figure 4(a), is an important factor in determining the radiation potential. The location of the panel captures the impact of geographical location, building orientation as well as the morphology of the surrounding of the building. The location can be expressed in terms of a cartesian coordinate of the centre point of the installation (i.e., x, y, z) or using ordinal data that represents the index of installation point, e.g., point "1" represents (x = 10, y = 15, z = 25). Both possibilities will be evaluated in this research; (2) *size of the panel*: the size of the panel, i.e., W and L in Figure 4(b), has an impact on the amount of solar radiation; (3) *orientation of the PV module*: this represents the tilt angle of the PV module in terms of θ; (4) *size and orientation of panels above the PV module*: the size and orientations of panels above any given PV module have an impact on the radiation potential of the PV module because of the shading effect they have on the panel. However, due to the impact of building orientation

on the incident angle of radiation, it can be imagined that it is not only the panel immediately above the studied PV module that may cast shade but also panels further away on left and right, as shown in Figure 5. Therefore, this study considers the size and orientation of panels above, top left and top right of the PV module. It should be highlighted that depending on the geographic location and orientation of the building, even panels further away might have a shading effect. However, in this research, it is assumed that the immediately adjacent panels capture the majority of the panel-induced shading effect. Therefore, the impact of panels further away can be ignored. In case any of these locations do not have a PV module, the W, L and θ are considered zero, as explained earlier in Figure 4(c).

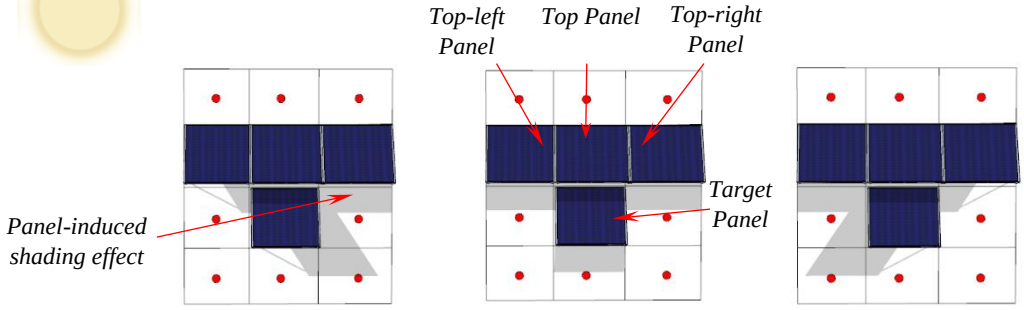


Figure 5 Shading effect of other surrounding PV modules

As for the labels (or output variables) of the machine learning model, the Annual Solar Radiation (ASR) can be used in terms of MWh. Ultimately, each data point in the dataset has a structure shown in Table 1. It should be highlighted that if the panel size is considered the same for all the panels, the pertinent variables can be ignored for the development of the machine learning model.

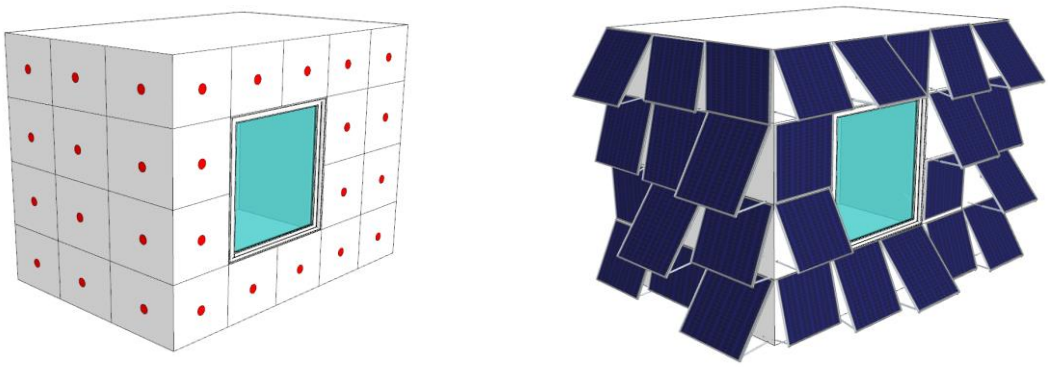
Table 1 Data structure for the development of the ML model of the PV panel layout design

Features (i.e., Input Variables)																**Label (i.e., Output)**
Point index	**Location**			**Size**		**Tilt**	**Top-left Panel**			**Top Panel**			**Top-right Panel**			**Annual Solar Radiation**
				Width	*Length*		*Width*	*Length*	*Tilt*	*Width*	*Length*	*Tilt*	*Width*	*Length*	*Tilt*	
i	X_i	Y_i	Z_i	W_i	L_i	θ_i	$W_{TL,i}$	$L_{TL,i}$	$\theta_{TL,i}$	$W_{T,i}$	$L_{T,i}$	$\theta_{T,i}$	$W_{TR,i}$	$L_{TR,i}$	$\theta_{TR,i}$	ASR_i

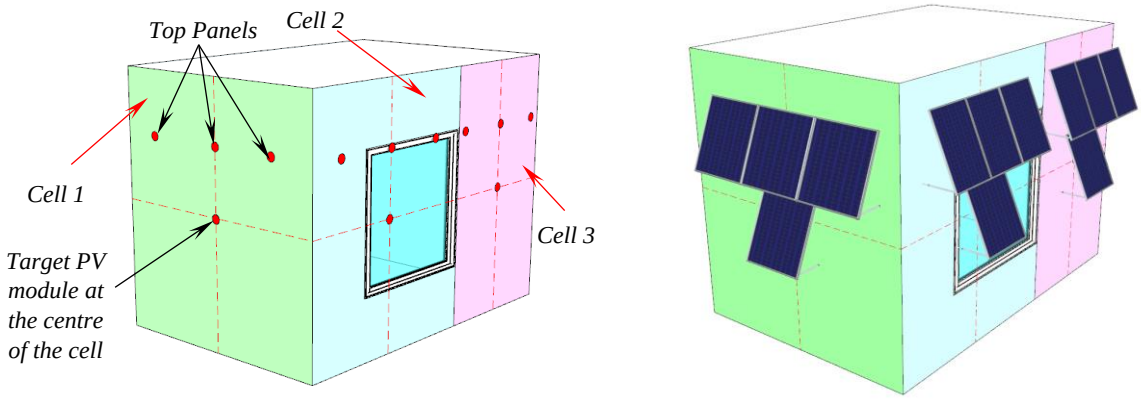
2.1.3. Development of different Scenarios

The common practice in the development of surrogate models is to generate a large number of random solutions that is well-balanced and distributed within the design space and use it as the dataset for the training. However, in the context of this research, a number of other scenarios can be envisioned to develop the ML model, mainly to simplify the dataset development process. In total, three different scenarios are considered in this research, as shown in Figure 6. In Scenario one, as shown in Figure 6(a), a number of entirely random solutions are generated using the parametric model explained above. This scenario is called random variation, and the same data structure presented in Table 1 is used to build the dataset.

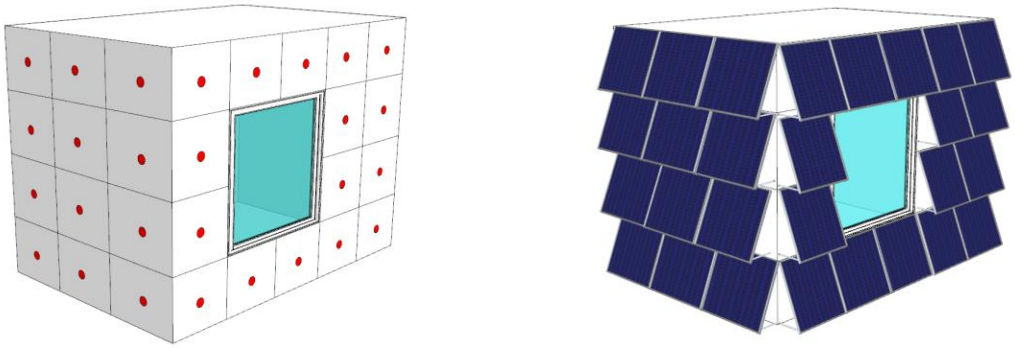
In Scenario 2, as shown in Figure 6(b), the design space is discretised into larger cells. It is assumed that the behaviour of PV modules in that portion of the façade can be represented by the four-panel layout shown in Figure 5. To this end, first different cells are formed on the façade of the building. Then, PV modules are placed at the centre of the cell as well as above, top left, and top right of the cell. Next, each panel is tilted between 0° to 90° with an increment of 10°. This procedure creates a combinatorial set consisting of 10,000 variations (i.e., 10 possible orientations for 4 panels generates 10^4 variations for the layout). This scenario is called Grid Variation. The structure presented in Table 1 is slightly modified for this scenario, in the sense that the grid label replaces the point label, and x, y, z of location is replaced with those of the centre of the cell.



(a) Random variation



(b) Grid variation



(c) Uniform Variation

Figure 6 Different scenarios studied in this research

This adjustment has a consequence in the Estimation phase and how the validation dataset needs to be prepared for this scenario. It will be explained in Section 2.1.4. It should be noted that since this grid is an approximation behaviour of the modules on the façade, the typical constraints, such as mechanical units, are disregarded in the placement of the representative modules in the grid. For instance, PV modules are placed in the centre of the window frame in the schematic Figure 6(b). This has no impact on the feasibility of the final solution because only the feasible PV modules belonging to each cell will be considered during the Estimation

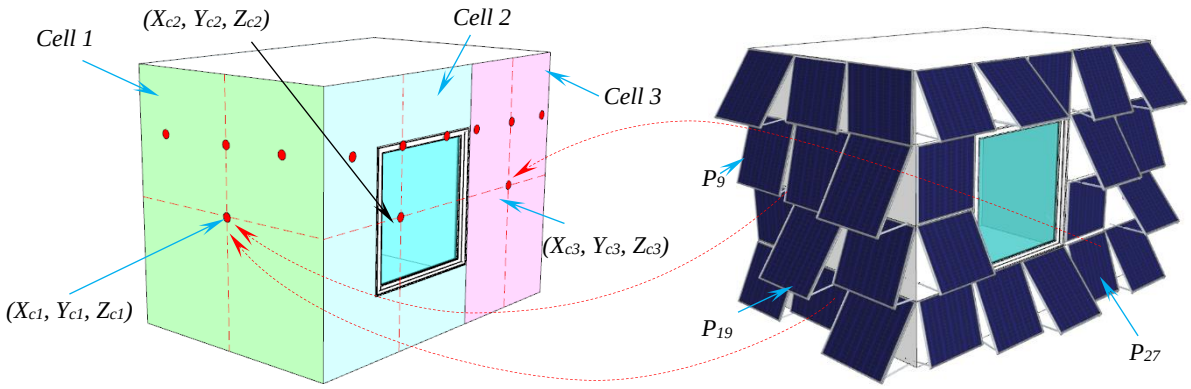
phase. It is hypothesised that the size of the cell has a negative impact on the accuracy of the surrogate model, meaning the larger the cell size, the lower the accuracy. This hypothesis will be tested in the case study by comparing the performances of ML models for three different sizes of cells.

In Scenario 3, shown in Figure 6(c), it is assumed that the relationship between the tilt angles of top panels and ASR of the target PV module is relatively linear and, therefore, can be approximated by linear functions. Thus, all panels are tilted in a uniform manner, meaning that all panels would always have the same angle. In this scenario, all panels are tilted from 0° to 90° using an increment of 5°. This scenario generates 19 different variations for the layout.

2.1.4. Splitting the datasets

Once datasets of different scenarios are built, they need to be split into training and validation datasets. While the training dataset is developed for each scenario, the validation dataset only contains solutions from Scenario 1, i.e., the Random Variation scenario. This is because regardless of how the ML model is developed, it is intended to approximate the solar radiation of any possible layout. Therefore, the model must be tested in cases where PV modules can be installed at any given location with any configuration.

However, some adjustments are required to prepare the validation dataset for the Grid Variation scenario, as shown in Figure 7. First, an extra feature needs to be added for each data point to represent cell to which each panel belongs. This can be done through simple point in polygon algorithm. Once the hosting cell of each panel is identified, the x, y, z of the PV module will be changed to the x, y, z of the cell. This is because the model in this scenario is built based on the cell coordinates.



Point	Features (i.e., Input Variables)																Label (i.e., Output)
	Cell index	Location			Size		Tilt	Top-left Panel			Top Panel			Top-right Panel			Annual Solar Radiation (MWh)
					Width (m)	*Length (m)*	*(°)*	*Width (m)*	*Length (m)*	*Tilt (°)*	*Width (m)*	*Length (m)*	*Tilt (°)*	*Width (m)*	*Length (m)*	*Tilt (°)*	
P_9	*1*	X_{c1}	Y_{c1}	Z_{c1}	1.5	1.5	20	0	0	0	1.5	1.5	35	1.5	1.5	10	0.657
P_{19}	*1*	X_{c1}	Y_{c1}	Z_{c1}	1.5	1.5	40	1.5	1.5	20	1.5	1.5	15	1.5	1.5	50	0.678
P_{27}	*3*	X_{c3}	Y_{c3}	Z_{c3}	1.5	1.5	16	0	0	0	1.5	1.5	0	1.5	1.5	48	0.632

Figure 7 Example of preparing the validation dataset for Grid Variation scenario

2.2. Development of Surrogate Model

The main steps of the proposed GA-based surrogate modelling method are shown in Figure 3. After datasets are generated, the ML model can be developed. In this research, a GA-based ML development approach is adopted to optimise the hyperparameters of the ML model. Hyperparameters refer to the basic configuration parameters of the ML model. The configuration of hyperparameters is shown to have a significant impact on the performance of the ML model (Han & Kim, 2019; Genuer et al., 2008; Wang et al., 2018). In this method, a population consisting of a random set of individual possible solutions is generated. Each solution to this population is called a *chromosome*. Each *chromosome* is divided into several parts, each consisting the hyperparameter geness of the ML method to be optimised. In each hyperparameter gene, each value in the given range will be used and paired with the value of other hyperparameters. This method will make multiple configurations to be evaluated. Figure 8 shows the structure of the chromosome in this method.

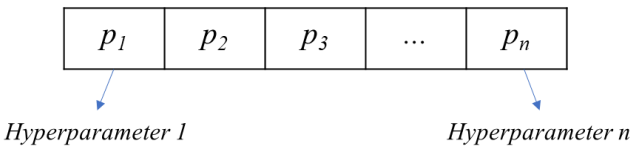


Figure 8 Structure of chromosome in the GA-based optimisation method

k-fold cross-validation method is used to train and test the ML model. Each *k*-1 subsamples from the training dataset will be used exactly once as the testing data. Then, as shown in Figure 9, the results from each iteration of the *k*−1 are averaged to estimate the final performance of each chromosome (Wainer & Cawley, 2017).

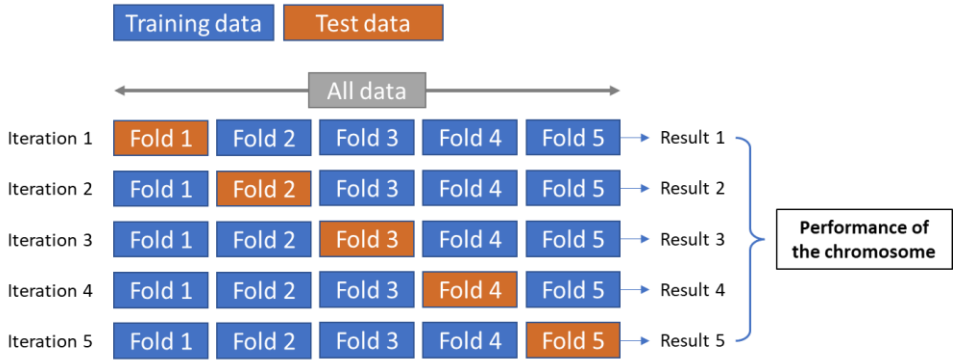


Figure 9 A general framework of k - fold cross-validation iteration

This research proposes the use of Random Forest (RF) as the ML method. RF is a popular and powerful supervised machine learning algorithm capable of performing both regression and classification tasks (Hasti et al., 2008). This method is selected mainly due to its demonstrated superiority in terms of handling the multi-dimensionality and imbalanced dataset (Brown & Mues, 2012; Ahmad et al., 2017; Langroodi et al., 2021). Besides, RF offers an approach to assess how important a variable is compared to the others, select the most important variables, and reduce dimensionality. Moreover, the parameters of RF are simple and computationally lighter than other machine learning methods, i.e., RF is computable even when it only has the number of decision trees (n_{tree}) and the number of input variables ($m_{variables}$) (Rodriguez-Giliano et al., 2014). Although the RF algorithm is well established (Breiman, 2001), a brief overview is presented for completeness.

RF is an ensemble method that combines several individual decision trees and forms the so-called forest. Every particular tree $\{h(\mathbf{x}, \Theta T), T = 1,2, \dots\}$ will be grown using the training set and the value of an independently-sampled random vector $\{\Theta T\}$, where this value is

distributed equally among each tree in that forest. The training data subsets for each tree are created through a procedure called bootstrapping. Bootstrapping creates training data by randomly resampling the original dataset without deleting the data selected from the input sample. This process makes the model more robust when facing slight variations in input data. Therefore, more excellent prediction stability can be achieved, and, at the same time, it increases prediction accuracy. When the target variable is continuous, RF uses Sum of errors or weighted Variance as the criterion for branching the tree at a node (Probst et al., 2019; Pedregosa, et al., 2011). Each selected feature will be calculated to explore the possible split point with minimum variance. In each possible split, the variance of each child node is individually calculated. Then, the variance of each split is computed as the weighted average variance of the child node. The one with the lowest variance value is selected as the best split (Sharma, 2020). At any step of the RF growth, the potential of the child nodes being a "leaf" must be controlled beforehand to define the end of the tree branch. A branch is considered a leaf when the information gain of the node is larger than any possible split to be considered. If it is still possible to split the node into two new nodes, it is not a leaf yet. This procedure is repeated until there are no more unbranched nodes and no more features left. The procedure of growing a regression tree is repeated for *T* trees in the forest.

After numerous trees are generated, the final prediction of RF uses the averaged value of the predictions from each tree (Breiman, 2001). Because RF is chosen in this research, the hyperparameters relevant to this study are shown in Table 2.

Table 2. Hyperparameters to be optimised

Hyperparameter	**Description**
n_estimators	The number of trees in the forest
max_depth	The maximum depth of each MTRT in the forest
max_features	The number of features to consider when looking for the best split
min_samples_leaf	The minimum number of samples required to be at a leaf node

The fitness function for the evaluation of the RF model is the accuracy of the prediction. In this research, the Mean Absolute Percentage Error ($MAPE$) is used to measure accuracy. Equation 1 presents $MAPE$ calculation method.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad \textit{Eq. 1}$$

Where:
n: *the number of samples*
y_i: *the actual value*
$\hat{y}_i$: *the predicted value*

The GA-based method stops if the improvement in the fitness function between two generations is smaller than a threshold or when the maximum generation number defined by the user has reached. The final RF model represents the optimum feature subset and hyperparameters. If the stopping criteria are not yet satisfied, another population of solutions will be generated through selection, crossover, mutation, and replacement.

2.3. Estimation

As the last phase, the ML model is used to estimate the solar radiation of the validation dataset. As stated in section 2.1.4, the validation dataset is a randomly selected subset of data from Scenario 1, i.e., Random Variation. Each includes a Cartesian coordinate (x, y, z) and the random tilt angle of the PV module, and also the random tilt angle of the top, top left, and top right panels, as shown in Table 1.

Finally, the (adjusted) validation dataset is fed into the ML model along with its best RF model developed in Phase 2, predicting the solar radiation amount. The accuracy performance of the prediction is then calculated using $MAPE$.

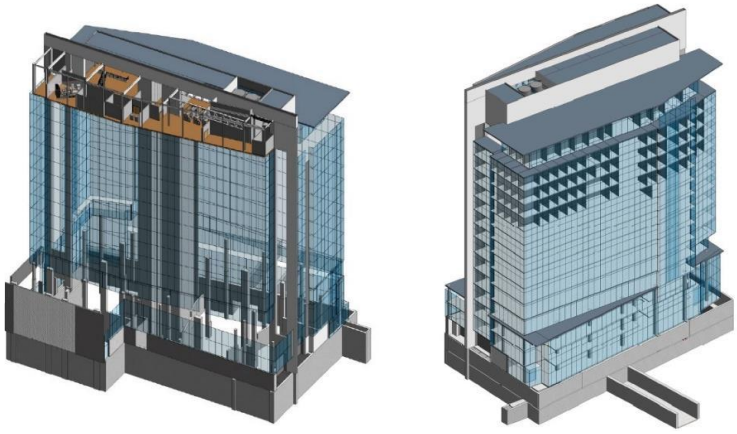
3. Case Study

A case study is conducted to test the performance and feasibility of the developed methodology. The case used in this research will be the John Molson School of Business (JMSB) building from Sir George William (SGW) campus of Concordia University in downtown Montreal Canada. This building stands 55 m tall with 15 storeys above the ground and an all-glass curtain wall as the façade (Concordia University, 2021). This is the same building used in the earlier research of the authors (Salimzadeh, et al., 2020). Having the same case study allows direct comparison of the results and, therefore, better identify the effect of the proposed method in improving the solar radiation simulation.

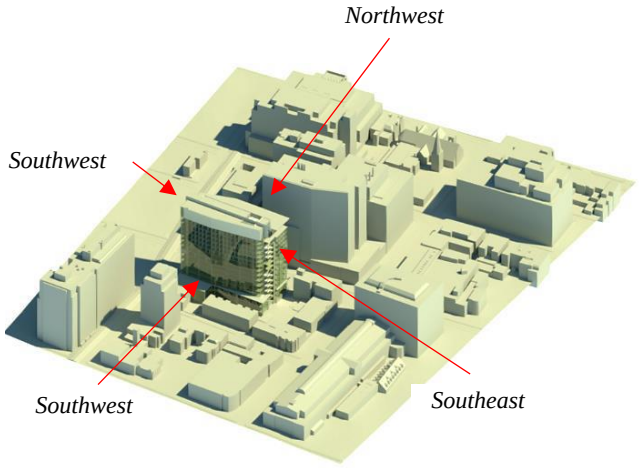
Although the process of preparing the 3D model for the solar radiation simulation was explained in the authors' previous work (Salimzadeh, et al., 2020), a brief overview is provided for completeness. As shown in Figure 10(a), Revit (Autodesk, 2021) was used to model the building in an object-oriented fashion, i.e., the BIM model of the building. This model was then integrated with the CityGML model of the surrounding buildings (City of Montreal, 2021), i.e., to consider the shadow effect of the neighbouring buildings in solar simulation in Revit environment, as shown in Figure 10(b). Inside Revit, Dynamo visual programming (Dynamo, 2021) was used to develop the PV layout parametric model. The implementation detail of this parametric model is provided in the authors' previous work (Salimzadeh, et al., 2020).

It should be noted that after the careful study of the surfaces in the JMSB building, it was discovered that the north-east façade of the building has a scant solar radiation potential because of the surrounding buildings. Therefore, this face of the façade was not considered for the installation of PV modules. The final configuration created a total of 1137 potential points for the installation of PV modules on the vertical surfaces of this study. It is essential to mention that in all these scenarios, the panel size was fixed at 2 × 1.5 m. Because of this

simplification, the features pertinent to the size of the panel were removed from the dataset, i.e., because they were uniform across all the layouts in the dataset.



(a)



(b)

Figure 10 (a) JMSB BIM model, (b) BIM Model and CityGML integration in Revit (Salimzadeh, et al., 2020)

As discussed in Section 2.1.3, three different scenarios were studied in this research. Table 3 shows the detail of the three scenarios and their configurations. As shown in this table, for the Random Variation scenario, 2200 random layouts were generated. For these random layouts, the option of not to put a PV module on a potential location was not considered because absence of PV module on a location result in the radiation of zero and also the impact of having no PV module on other surrounding panels is equal to having a panel with zero tilt angle. Therefore, the random solutions only include the random variation of PV modules tilt angles in the range of 0° to 90°. Of the 2200 generated solutions, 200 were set aside as the validation dataset used

to validate all three scenarios' performance. The validation dataset, consisting of 200 random layouts, was used to estimate the performance of the surrogate model. For the grid approach, as explained in Section 2.1.4, the location coordinates of panels were adjusted.

Four different configurations of this scenario were considered to test the impact of the dataset's size on the surrogate model's performance, using 500, 1000, 1500, and 2000 of the random solutions as the training dataset.

Table 3 The configurations of different scenarios used in this research

Scenario	**Training dataset**	**Validation dataset**
Scenario 1: Random Variation	500 layouts	200 random layouts
	1000 layouts	
	1500 layouts	
	2000 layouts	
Scenario 2: Grid Variation	Large grid: 26 cells	
	Medium grid: 37 Cells	
	Small Grid: 56 Cells	
Scenario 3: Uniform Variation	19 uniform layouts	

In Scenario 2, three different sizes of grids were considered, as shown in Figure 11. These grid sizes were first built based on the surface given by using Dynamo nodes of $Topology.Vertices$ and $Vertex.PointGeometry$. These nodes allow user to take a list of geometry surfaces containing a layout and export the corner point coordinates of those surfaces. The first surface export gives the corner points of the cell as used as a whole for the large grid, as given in Figure 11(a). For the medium grid, the sizes of the cells are reduced to 50%, except those containing less than 50 panels on the comprehensive case. Meanwhile, for the small grid, the size of the cells is reduced to 33% from the initial grid, except for those cells that contain less than 25 comprehensive panels. These thresholds created 26, 37, and 56 cells for large, medium and small grids, respectively. As explained in Section 2.1.3, each cell contained one PV module in the centre and three PV modules on the top, creating a 4-panel system that is expected to approximate the behaviour of PV modules in each cell.

Scenario 3, as mentioned in Section 2.1.3, includes uniformly tilted panels with the tilt angle varying between 0° to 90° with the steps of 5°, generating an overall of 19 distinct solutions for the ML training.

It should be noted that as mentioned in Section 2.1.2, each configuration of the training dataset was used to develop two surrogate model, once with the cartesian coordinate and once with the index representing the specific panel location, as Shown in Table 1.

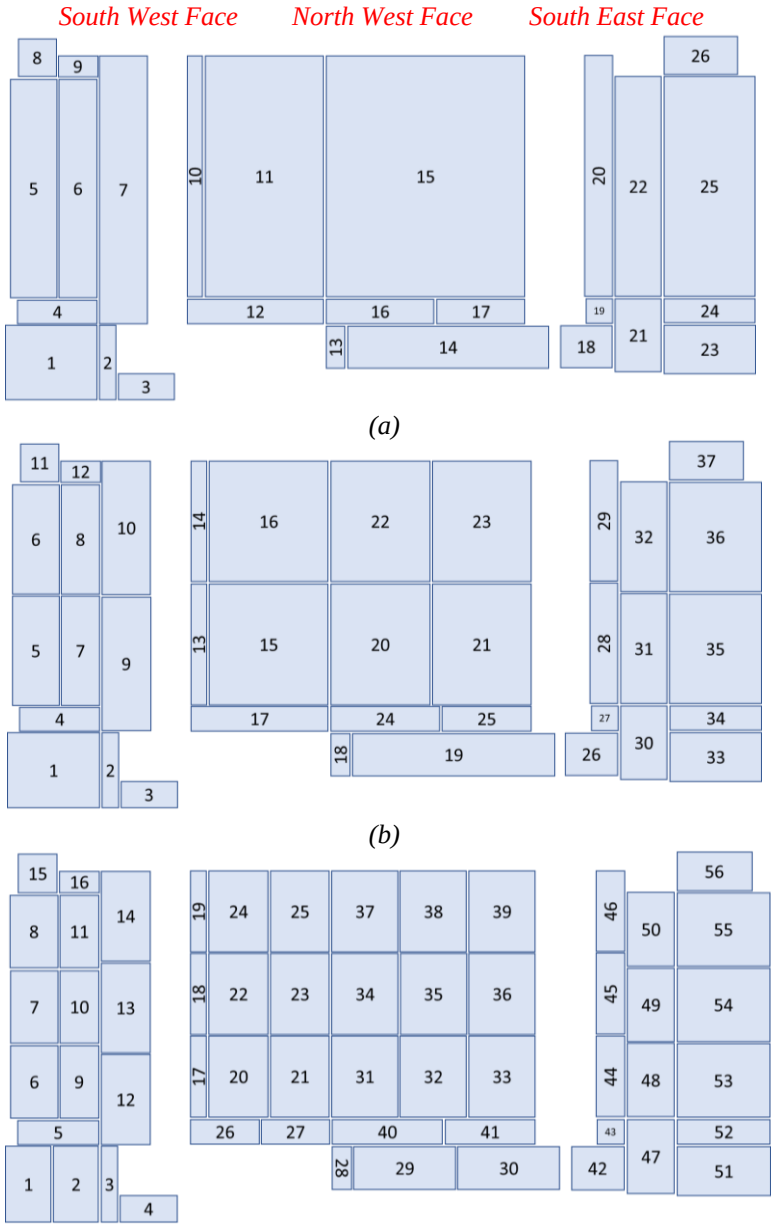


Figure 11 (a) Large, (b) Medium, and (c) small grid layout

As explained in Section 2.2, a GA-based RF model is proposed in this research. Table 4 presents the ranges of RF hyperparameters explored in this research. Concerning the minimum number of samples required for a node to be a leaf, this value is not recommended to use the default value of 1 because it can cause overfitting in cases where the training dataset is very large (Mantovani, et al., 2018). The "auto" in max_feature means that all features are considered when looking for the best split. For those with n values means that there are n features considered in each split.

Table 4 Range of hyperparameters to be optimised

Hyperparameter	Range
n_estimators	100, 250, 500, 750, 1000
max_depth	7, 15, 25, 50
max_features	'auto', 2, 3, 4
min_samples_leaf	10, 25, 50

Also, the configuration of the GA used for the optimisation of RF is presented in Table 5. The initial generation number was set to 20, which means that GA ran 20 iterations to find the near-optimum RF model. In each generation, 50 individuals were generated. For the cross-validation, a 5-fold cross-validation structure was used, with $MAPE$ being the minimisation objective function.

Table 5 Configuration of GA algorithm

GA Parameters	Description	Value
Generations	Number of iterations to run the pipeline optimisation process.	20
Population_size	The number of individuals that will be evaluated and selected.	50
Scoring	Function used to evaluate the quality of a given pipeline for the regression problem.	$MAPE$
cv	Number of folds in cross-validation strategy to be used when evaluating pipelines.	5

4. Results and Discussion

As stated in Section 2.3, the accuracy performance of the developed ML model is assessed using $MAPE$. However, in regression problems, researchers usually also use R^2 to demonstrate the accuracy and performance of the ML model's prediction (Breiman, 2001; Rodriguez-Galiano et al., 2014; Biau & Scornet, 2016). Therefore, to confirm the accuracy performance of each model, this research shall use two estimations, i.e., $MAPE$ and R^2. Unlike $MAPE$, if the performance's value of R^2 is closer to 1, it means that the accuracy is higher.

Table 6 presents the optimal configuration of RF hyperparameters for each scenario. This table allows acknowledging the result of hyperparameter optimisation if using different approaches. As shown in this table, optimum hyperparameters are relatively consistent across different scenarios. Interestingly, the maximum performance was achieved with the smallest number of trees, i.e., estimators, in the RF. Regarding the considered features that gives the best performance, most of the optimisation returns "auto" which means those scenarios consider all features when looking for the best split and no features are excluded. However, two scenarios, i.e., Random variation with 1500 layouts and 2000 layouts, showed that the number of considered features are only four. Both applies to when the coordinates of the points are used in the training.

Table 6 Results of the optimal hyperparameters

Scenario		**n_estimators**		**max_depth**		**max_features**		**min_samples_leaf**	
		Coordinate	*Index*	*Coordinate*	*Index*	*Coordinate*	*Index*	*Coordinate*	*Index*
Random Variation	*500 layouts*	*100*	*100*	*25*	*25*	*"auto"*	*"auto"*	*10*	*10*
	1000 layouts	*100*	*100*	*25*	*25*	*"auto"*	*"auto"*	*10*	*25*
	1500 layouts	*100*	*100*	*25*	*50*	*4*	*"auto"*	*10*	*10*
	2000 layouts	*100*	*100*	*25*	*50*	*4*	*"auto"*	*10*	*10*
Grid Variation	*Large grid*	*100*	*100*	*25*	*15*	*"auto"*	*"auto"*	*10*	*50*
	Medium grid	*100*	*100*	*25*	*7*	*"auto"*	*"auto"*	*50*	*50*
	Small grid	*100*	*100*	*25*	*25*	*"auto"*	*"auto"*	*10*	*10*
Uniform Variation		*100*	*100*	*25*	*25*	*"auto"*	*"auto"*	*10*	*10*

Table 7 shows the result of feature importance of the developed models. This table gives which features are the best and which has little to almost no contribution with the prediction of the data in each scenario. The most important feature for all of the models are related to the location of the point, but different location height, i.e., the z coordinate, gives less impact to the prediction. Even for some models, the different location height is less important than the tilt angle of the panel itself. Also, for most scenarios, different input of the surrounding panels' tilt angle are the least important in predicting the solar panel amount.

Table 7 Result of feature importance

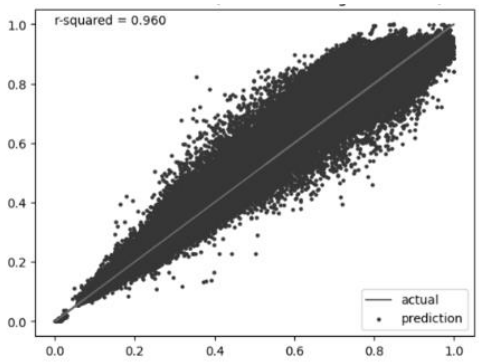
Type of Scenario Approaches			**Location**				**Tilt**			
			Index	*X*	*Y*	*Z*	*Panel*	*Top Panel*	*Top-right Panel*	*Top-left Panel*
Random	**500 layouts**	*Coordinate*	-	32.79%	**48.39%**	8.21%	8.34%	1.25%	**0.49%**	0.52%
		Index	**75.13%**	-	-	-	8.50%	**1.92%**	7.01%	7.44%
	1000 layouts	*Coordinate*	-	32.42%	**48.83%**	8.23%	8.29%	1.25%	**0.47%**	0.52%
		Index	**75.84%**	-	-	-	8.26%	**1.70%**	6.93%	7.27%
	1500 layouts	*Coordinate*	-	**41.30%**	37.27%	9.27%	8.42%	1.31%	1.31%	**1.12%**
		Index	**76.50%**	-	-	-	8.05%	**1.84%**	6.62%	6.99%
	2000 layouts	*Coordinate*	-	**41.85%**	36.83%	9.39%	8.41%	1.25%	1.16%	**1.11%**
		Index	**76.55%**	-	-	-	8.00%	**1.84%**	6.62%	6.98%
Grid	**Small**	*Coordinate*	-	28.89%	**55.52%**	3.04%	11.73%	0.44%	0.32%	**0.06%**
		Index	**87.01%**	-	-	-	12.18%	0.44%	0.32%	**0.06%**
	Medium	*Coordinate*	-	27.29%	**55.96%**	4.41%	11.58%	0.36%	0.37%	**0.04%**
		Index	**87.26%**	-	-	-	11.99%	0.35%	0.36%	**0.03%**
	Large	*Coordinate*	-	24.34%	**55.46%**	7.30%	12.09%	0.37%	0.42%	**0.04%**
		Index	**87.88%**	-	-	-	11.36%	0.34%	0.40%	**0.03%**
Uniform		*Coordinate*	-	36.44%	**48.22%**	6.61%	5.84%	1.44%	**0.36%**	1.08%
		Index	**68.32%**	-	-	-	4.40%	**3.66%**	12.22%	11.40%

Table 8 presents the results of the estimation of various surrogate models developed in this research. Also, Figures 12 to 14 show the regression plots of different surrogate models. Since all the models use the same set for the validation dataset, the regression plots for generated solar radiation are normalised.

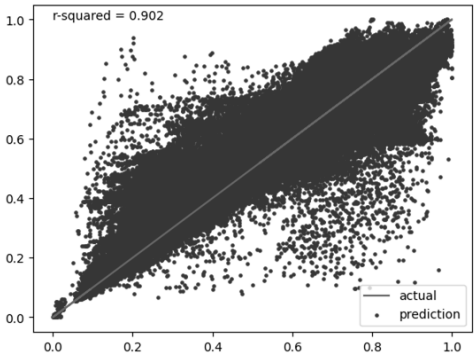
Table 8 Performance of different Surrogate models

Type of Approaches		**Coordinate**		**Index**	
		R-squared	*MAPE (%)*	*R-squared*	*MAPE (%)*
Random Variation	*500 layouts*	*0.96*	*6.70*	*0.90*	*10.71*
	1000 layouts	*0.96*	*6.61*	*0.90*	*10.64*
	1500 layouts	*0.96*	*6.57*	*0.95*	*7.27*
	2000 layouts	*0.96*	*6.55*	*0.95*	*7.18*
Grid Variation	*Large grid*	*0.53*	*35.54*	*0.54*	*35.04*
	Medium grid	*0.55*	*34.70*	*0.55*	*34.56*
	Small grid	*0.55*	*34.67*	*0.55*	*34.64*
Uniform Variation		*-1.41*	*61.98*	*-1.35*	*81.52*

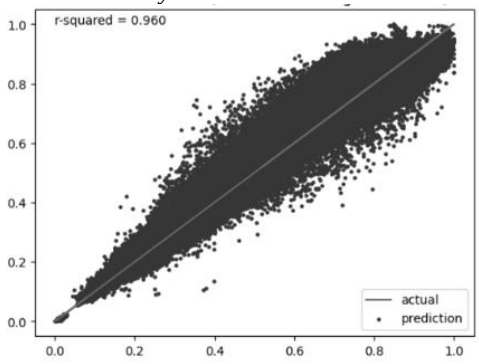
From three different scenarios, all configurations of the Random variation scenario perform decisively better than other scenarios. This result is interesting because the dataset size for this scenario is orders of magnitude smaller than that of the Grid variation scenario, as explained in Section 2.1.3. In the Random variation scenario, it seems that the size of the dataset has little impact on the performance of the ML model, more so when the coordinates of the points are used for training. When training with the index of points, then the size of the dataset became more influential, where the larger dataset performed better in terms of both R^2 and MAPE. While the use of index instead of coordinate had a negative impact on the accuracy of the models in Random variation and Uniform variation scenarios, it had a minimal positive impact in the Grid variation scenario. Looking at the Random variation scenario, the negative impact of using an index instead of coordinates became smaller with the increase in the size of the training dataset, showing that index-based training is more sensitive to the size of the dataset. The fact that even the smallest dataset in the Random variation dataset had a high performance is promising because it indicates that even with a small number of randomly generated layouts, a reliable and accurate surrogate model can be developed.



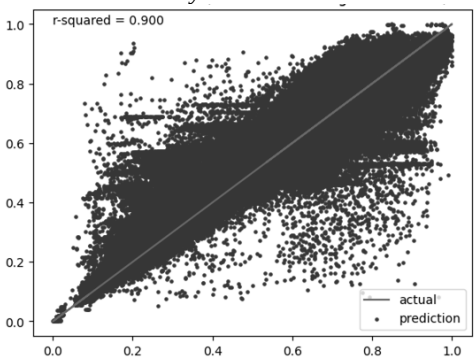
500 layouts with Coordinates



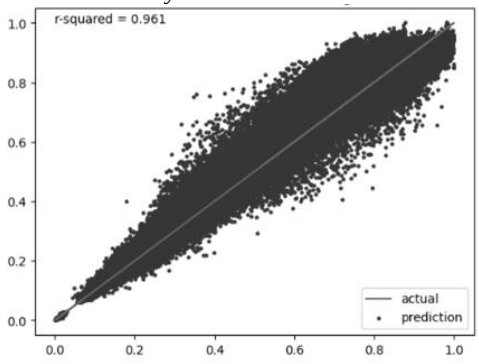
500 layouts with Index



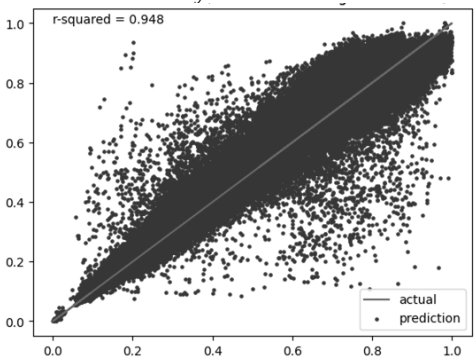
1000 layouts with Coordinates



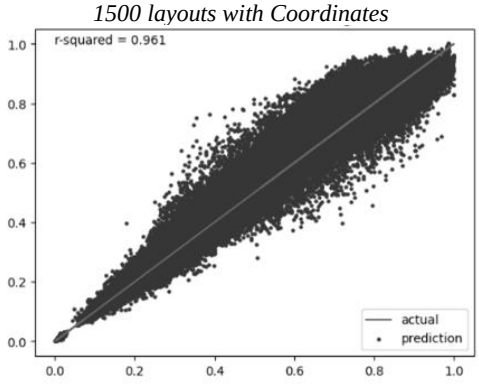
1000 layouts with Index



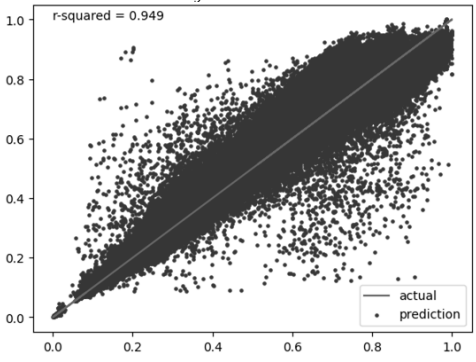
1500 layouts with Coordinates



1500 layouts with Index

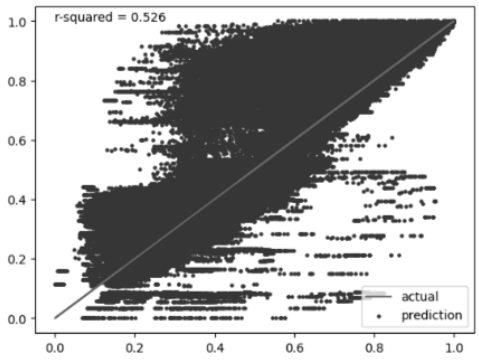


2000 layouts with Coordinates

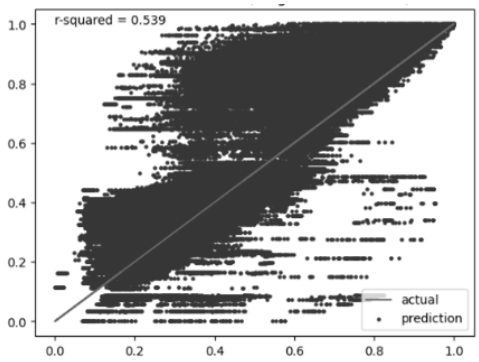


2000 layouts with Index

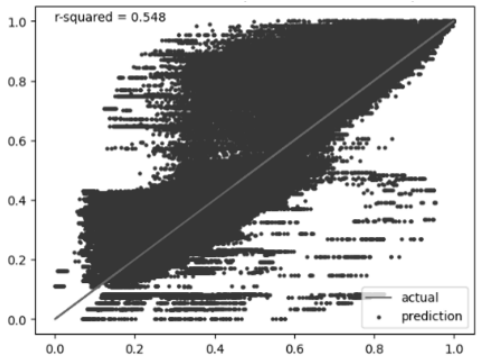
Figure 12 Regression plots of different configurations of Random variation scenario



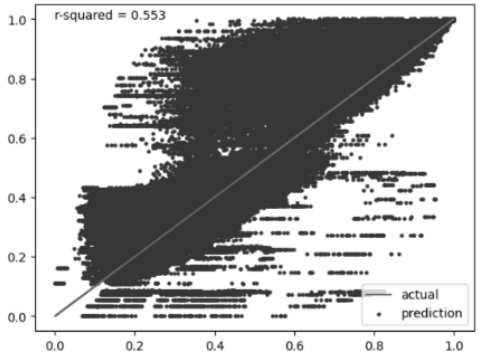
Large grid with Coordinates



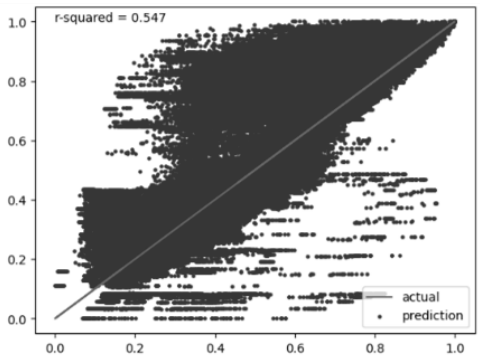
Large grid with Index



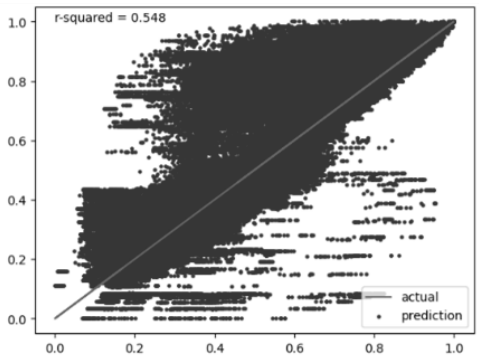
Medium grid with Coordinates



Medium grid with Index

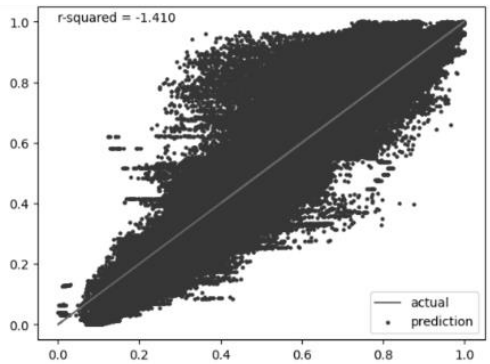


Small grid with Coordinates

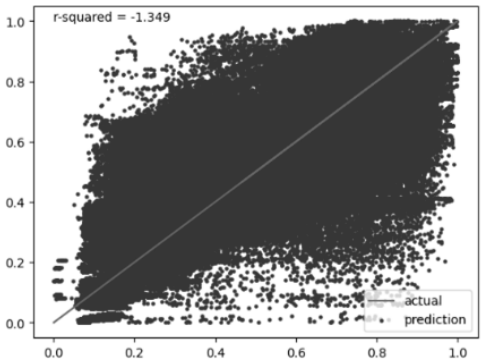


Small grid with Index

Figure 13 Regression plots of different configurations of Grid variation scenario



Uniform variation with Coordinates



Uniform variation with Index

Figure 14 Regression plots of Uniform variation scenario

Despite having massive training datasets, all configurations of the Grid variation scenario showed low accuracy. Nevertheless, the increase in the number of cells, i.e., smaller grid, proved to have a slight but positive effect on the model's performance. A few points in the regression plots with the greatest deviation were analysed to understand better why this scenario did not perform well. It was observed that these deviations belong to locations on the façade which, based on the grid-based approximation, were supposed to generate a large amount of radiation because the surrounding panels had little to no tilt angle. However, when projected to the actual layout, it was observed that these panels received considerable shadow effect from the building façade or surrounding buildings. Although the result of this scenario is not favourable, it provides the valuable insight that the simulation of solar radiation is very sensitive to the location and configuration of the panels, so that an accurate approximation cannot be made by applying zoning on the façade. Nevertheless, as shown in the previous scenario, it is evident that by simulating a random solution sample for each potential installation point, a high-accuracy prediction can be made on the solar radiation potential.

The above observation about the complexity of the solar radiation simulation is further corroborated by the very low performance of the Uniform variation scenario. It is demonstrated that an accurate prediction can be made by using a small dataset consisting of a Uniform variation of tilt angles. This shows the high sensitivity of the solar radiation simulation to the shadow effect of the surrounding panels.

In terms of the simulation time needed, generating the radiation amount of one random solution configuration in Dynamo requires approximately 19 seconds. Using the RF model, 200 random solutions from the validation dataset requires 236 seconds to export the amount of solar radiation obtained. This means the RF model takes 1.18 seconds to generate one solution. Therefore, using the RF model is 16 times faster compared to using the simulation-based parametric model.

5. Conclusion

This research investigated the performance of a surrogate modelling approach for the simulation of solar radiation potential on the vertical surfaces of tall buildings. Surrogate modelling was used to approximate the input-output behaviour of the existing simulation model. The RF machine learning approach was used in investigating three different scenarios, namely (1) Random variation, (2) Grid variation, and (3) Uniform variation. GA was used to optimise the hyperparameters of the RF model. A case study was performed to investigate the performance of surrogate models. The case study used a building in Sir George William (SGW) campus of Concordia University in downtown Montreal Canada.

It was demonstrated that, in general, surrogate modelling has a great potential to accurately approximate the simulation of solar radiation on the vertical surfaces of tall buildings. It was shown that accuracy of up to 94% could be achieved even by only using a small sample of data, i.e., 500 random layouts. In fact, the surrogate model is capable to give the result of solar radiation amount by 16 times faster compared to the existing simulation model. This development can help to tremendously reduce the computational intensity of optimisation-based PV model layout design. However, it was observed that the best approach to develop the surrogate model is to use a number of random layout designs rather than more guided strategies, such as grid-based approximation or uniform variation of tilt angles. This attests to the fact that while surrogate modelling is very promising and applicable, the solar radiation simulation is very complex and too sensitive to the location and shadow effects. Because of this sensitivity, simplification cannot be made to approximate the solar radiation potential. Nevertheless, even a small sample of random design layout that captures the diversity of panel configurations for all the potential locations can be used to predict solar radiation potential accurately.

However, there are a few limitations to this research. First, this study only considers using Random Forest for developing the ML models. It is possible to use other types of ML methods, such as Neural-Network-based methods. Also, the PV panel size in this study only was fixed at 2 x 1.5 m. It is also possible to add other possible panel sizes to see how different sizes affect ML models' prediction and performance. Finally, although the developed surrogate model could be easily used to optimise the PV layout, i.e., to perform a generative design, it was out of the scope of this research. In the future, the authors intend to perform generative design based on this surrogate model and then compare the results of the design optimisation with that of simulation-based optimisation to investigate to what extent the use of the surrogate model can contribute to finding better layout design and faster.

6. References

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77-89.

Al-Janahi, S. A., Ellabban, O., & Al-Ghamdi, S. G. (2020). A Novel BIPV reconfiguration algorithm for maximum power generation under partial shading. *Energies, 13(17)*, 44-70.

Arisha, A., & Abo-Hamad, W. (2010). Simulation Optimisation Methods in SupplyChain Applications: A Review. *Irish Journal of Management 30 (2)*, 95-124.

Autodesk. (2021). *Revit*. Retrieved from Autodesk: http://www.autodesk.com/education/free-software/revit

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST 25*, 197-227.

Bornatico, R., Hüssy, J., Witzig, A., & Guzzella, L. (2013). Surrogate modeling for the fast optimization of energy systems. *Energy, Vol. 57*, 653-662.

Breiman, L. (2001). Random Forests. *Machine Learning Volume 45*, 5-32.

Brown. (2016, March). *How to Run a Solar Radiation Analysis in Revit*. Retrieved from Dylan Brown Designs: http://dylanbrowndesigns.com/tutorials/how-to-run-a-solar-radiation-analysis-in-revit/

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

Carneiro, C., Morello, E., Desthieux, G., Golay, F., & . (2010). Urban environment quality indicators: application to solar radiation and morphological analysis on built area. *3rd WSEAS international conference on Visualization, imaging and simulation* (pp. 141-148). World Scientific and Engineering Academy and Society (WSEAS).

Catita, C., Redweik, P., & Pereira, J. (2014). Extending solar potential analysis in buildings to vertical facades. *Computers & Geosciences Volume 66*.

Chow, A., Fung, A. S., & Li, S. (2014). GIS modeling of solar neighborhood potential at a fine spatiotemporal resolution. *Buildings, 4(2)*, 195-206.

City of Montreal. (2021). *Maquette numérique (Bâtiments CityGML LOD2 avec textures)*. Retrieved from Portail des données ouvertes: http://donnees.ville.montreal.qc.ca/

Concordia University. (2021). *Molson Building features*. Retrieved from Concordia University: https://www.concordia.ca/maps/buildings/mb/highlights-features.html

Dynamo. (2021). *Open source graphical programming for design*. Retrieved from http://dynamobim.org/

Elfaki, A. O., Alatawi, S., & Abushandi, E. (2014). Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey. *Advances in Civil Engineering*, 1-11.

Esclapés, J., Ferreiro, I., Piera, J., Teller, J., & . (2014). A method to evaluate the adaptability of photovoltaic energy on urban façades. *Solar Energy Vol 105*, 414-427.

Freitas, S., Serra, F., & Brito, M. C. (2015). PV layout optimization: String tiling using a multiobjective genetic algorithm. *Energy, 118*, 562-574.

Genuer, R., Poggi, J. M., & Tuleau, C. (2008). *andom Forests: some methodological insights*. Retrieved from ArXiv Preprint: https://arxiv.org/abs/0811.3619v1

Gibson, E. (2017, August). *CF Moller cover Copenhagen school in 20,000 solar panels*. Retrieved from de zeen: https://www.dezeen.com/2017/08/23/copenhagen-international-school-c-f-moller-architects-12000-solar-panels-denmark/

Godoy-Shimizu, D., Steadman, P., Hamilton, I., Donn, M., Evans, S., Moreno, G., & Shayesteh, H. (2018). Energy use and height in office buildings. *Building Research & Information*, 845-863.

Gooding, J., Edwards, H., Giesekam, J., & Crook, R. (2013). Solar City Indicator: A methodology to predict city level PV installed capacity by combining physical capacity and socioeconomic factors. *Solar Energy, 95*, 325-335.

Gurupira, T., & Rix, A. J. (2017). Constrained optimisation of photovoltaic (PV) module layouts. *IEEE Africon*, (pp. 1179-1184).

Han, S., & Kim, H. (2019). On the Optimal Size of Candidate Feature Set in Random forest. *Applied Sciences 2019, Vol 9(5)*, 898.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). Random Forest. In *The Elements of Statistical Learning* (pp. 587-604). New York: Springer.

Hwang, T., Kang, S., & Kim, J. T. (2012). Optimization of the building integrated photovoltaic system in office buildings—Focus on the orientation, inclined angle and installed area. *Energy and Buildings, 46*, 92-104.

IEA, I. E. (2018). *Renewables*. Retrieved from IEA: https://www.iea.org/topics/renewables/subtopics/solar/

Jelle, B. P., Breivik, C., & Røkenes, H. D. (2012). Building integrated photovoltaic products: A state-of-the-art review and future research opportunities. *Solar Energy Materials and Solar Cells, 100*, 69-96.

Jochem, A., Höfle, B., Rutzinger, M., & Pfeifer, N. (2009). Automatic roof plane detection and analysis in airborne lidar point clouds for solar potential assessment. *Sensors, 9(7)*, 5241-5262.

Kåberger, T. (2018). Progress of renewable electricity replacing fossil fuels. *Global Energy Interconnection, Volume 1, Issue 1,*, 48-52.

Karnon, J., Stahl, J., Brennan, A., Caro, J. J., Mar, J., & Möller, J. (2012). Modeling using Discrete Event Simulation: A Report of the ISPOR-SMDM. *Value in Health 15*, 821–827.

Khatib, H. (2012). IEA World Energy Outlook 2011—A comment. In A. Midttun, & A. Martinelli, *Energy Policy Special Section: Frontiers of Sustainability Volume 48* (pp. 737-743). World Energy Council.

Koo, C., Hong, T., Lee, M., & Kim, J. (2016). An integrated multi-objective optimization model for determining the optimal solution in implementing the rooftop photovoltaic system. *Renewable and Sustainable Energy Reviews, 57*, 822-837.

Kucuksari, S., Khaleghi, A. M., Hamidi, M., Zhang, Y., Szidarovszky, F., Bayraksan, G., & Son, Y. J. (2014). An Integrated GIS, optimization and simulation framework for optimal PV size and location in campus area environments. *Applied Energy, 113*, 1601-1613.

Kumar, L., Skidmore, A. K., & Knowles, E. (1997). Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science, 11(5)*, 475-497.

Langroodi, A. K., Vahdatikhaki, F., & Doree, A. (2021). Activity recognition of construction equipment using fractional random forest. *Automation in Construction Vol 122*, 1-17.

Liang, J., Gong, J., Li, W., & Ibrahim, A. N. (2014). A visualization-oriented 3D method for efficient computation of urban solar radiation based on 3D–2D surface mapping. *International Journal of Geographical Information Science 28(4)*, 780-798.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv:1812.02207v2*.

Mardaljevic, J. (2000). Simulation of annual daylighting profiles for internal illuminance. *International Journal of Lighting Research and Technology 32(3)*, 111-118.

Marszal, A. J., Bourrelle, J. S., Musall, E., Voss, K., Sartori, I., & Napolitano, A. (2011). Zero Energy Building–A review of definitions and calculation methodologies. *Energy and buildings, 43(4)*, 971-979.

Martín, A. M., Domínguez, J., & Amador, J. (2015). Applying LiDAR datasets and GIS based model to evaluate solar potential over roofs: A review. *AIMS Energy, 11(2)*, 326-343.

Middelhauve, L., Baldi, F., Stadler, P., & Maréchal, F. (2021). Grid-Aware Layout of Photovoltaic Panels in Sustainable Building Energy Systems. *Frontiers in Energy Research Vol 8*, 1-19.

Nguyen, A.-T., Reiter, S., & Rigo, P. (2014). A review on simulation-based optimization methods applied to building performance analysis. *Applied Energy Vol 113*, 1043-1058.

Ning, G., Kan, H., Zhifeng, Q., Weihua, G., & Geert, D. (2018). e-BIM: a BIM-centric design and analysis software for Building Integrated Photovoltaics. *Automation in Construction, 87*, 127-137.

Numan, A. H., Dawood, Z. S., & Hussein, H. A. (2020). Theoretical and experimental analysis of photovoltaic module characteristics under different partial shading conditions. *International Journal of Power Electronics and Drive System Vol 11 (3)*, 1508-1518.

Paulescu, M., Paulescu, E., Gravila, P., & Badescu, V. (2012). *Weather modeling and forecasting of PV systems operation.* Springer Science & Business Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *JMLR 12 (85)*, 2825−2830.

Perera, A. T., Wickramasinghe, P. U., Nik, V. M., & Scartezzini, J. L. (2019). Machine learning methods to assist energy system optimization. *Applied Energy, Vol 243*, 191-205.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3)*, 1-15.

Raugei, M., & Frankl, P. (2009). Life cycle impacts and costs of photovoltaic systems: current state of the art and future outlooks. *Energy, 34(3)*, 392-399.

Resch, E., Bohne, R. A., Kvamsdal, T., & Lohne, J. (2016). Impact of urban density and building height on energy use in cities. *Energy Procedia 96*, 800 – 814.

Robinson, D., & Stone, A. (2004). Irradiation modelling made simple: the cumulative sky approach and its applications. *In PLEA conference*, (pp. 19-22).

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of The Total Environment*, 476-477.

Salimzadeh, N. (2021). 'Optimization of PV Modules Layout on High-rise Building Skins Using a BIM-based Generative Design Approach'. PhD thesis, Concordia University, Montreal.

Salimzadeh, N., Vahdatikhaki, F., Hammad, A., ., & . (2020). Parametric modeling and surface-specific sensitivity analysis of PV module layout on building skin using BIM. *Energy & Buildings 216*, 1-14.

Sharma, A. (2020, June 30). *4 Simple Ways to Split a Decision Tree in Machine Learning*. Retrieved from Analytic Vidhya: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

Smith, A., & Gill, G. (2014). *FKI Tower*. Retrieved from Smith Gill: http://smithgill.com/work/fki/

Thrampoulidis, E., Mavromatidis, G., Lucchi, A., & Orehounig, K. (2021). A machine learning-based surrogate model to approximate optimal. *Applied Energy 281*, 1-20.

Wainer, J., & Cawley, G. (2017). Empirical Evaluation of Resampling Procedures for. *Journal of Machine Learning Research 18* , 1-35.

Wang, Q., Nguyen, T.-T., Huang, J. Z., & Nguyen, T. T. (2018). An efficient random forests algorithm for high dimensional data classification. *Advances in Data Analysis and Classification, 12(4)*, 953-972.

Wang, Z., & Schafer, B. C. (2020). Machine leaming to set meta-heuristic specific parameters for high-level synthesis design space exploration. *57th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). San Francisco: Institute of Electrical and Electronics Engineers Inc.

Ward, G. J. (1994). The RADIANCE lighting simulation and rendering system. *The 21st annual conference on Computer graphics and interactive techniques* (pp. 459-472). ACM.

Xu, W., Lam, K. P., & Karaguzel, O. T. (2017). Using an adaptive meta-model evolutionary algorithm for mixed-integer type building design optimization. *15th International IBPSA Conference*, (pp. 1849 - 1858). San Francisco.

Yüksek, I., & Karadayi, T. T. (2017). Energy-Efficient Building Design in the Context of Building Life Cycle. In *Energy Efficient Buildings* (pp. 93-123). Intech Open.