SimAUD 2021
SYMPOSIUM ON SIMULATION FOR ARCHITECTURE + URBAN DESIGN

Predicting Incident Solar Radiation on Building's Envelope Using Machine Learning

Ammar Alammar[12], Wassim Jabi[1], Simon Lannon[1]

[1]Cardiff University
Cardiff, United Kingdom
{alammara, jabiw, lannon}@cardiff.ac.uk

[2]King Saud University
Riyadh, Saudi Arabia
aammar@ksu.edu.sa

ABSTRACT

The assessment of the impact of solar radiation on building envelopes has typically been achieved by using simulation software, which is time consuming and requires advanced computational knowledge. Given the increased complexity of large scale-projects and the demand for performative buildings, new innovative methods are required to assess the design efficiently. In this paper, we present an alternative and innovative approach to assessing solar radiation intensity on an office building envelope using two machine-learning (ML) models: Artificial Neural Network (ANN) and Decision Tree (DT). The experimental workflow of this paper consists of two stages. In the first stage, a generative parametric office tower and its urban context were designed and simulated using Grasshopper software to create a large synthetic dataset of the solar radiation that strikes the office room envelope with several types of analyses. In the second stage, the generated datasets were imported into two ML algorithms (ANN and DT) to create a model for training and testing. The comparison of these two ML models proved that input data types have a significant impact on the accuracy of the prediction and model selection. DT was found to be more accurate than ANN because the data is mostly categorical, which is the most suitable learning background for DT algorithms.

Author Keywords
Machine Learning, Solar Radiation, Decision Tree, Neural Network

ACM Classification Keywords
I.6.1 SIMULATION AND MODELING

1 INTRODUCTION

The challenge of creating a sustainable environment requires the incorporation of building performance evaluation in the early stages of the design process [12]. To support the decision-making process during the conceptual phase of the design, the initial concept of the design needs to be evaluated considering different analysis studies to achieve high-performance buildings. Solar radiation is one of the main environmental studies that need to be investigated initially to decide on the form, orientation, and envelope of high-rise buildings [2]. In hot, arid climates, high-rise building surfaces are exposed to intense solar energy with direct and diffused radiation throughout the year. In addition, a large amount of solar radiation strikes these surfaces during the summer season [45]. If vertical services and glazed façades are not evaluated thoroughly in the first stage of the design, solar thermal energy will penetrate into the interior spaces, thus affecting both the human comfort level and energy consumption [31].

The machine learning (ML) approach promises greater efficiency in the evaluation of building performance than does conventional simulation [8]. Moreover, ML has major benefits including the reduction of computational time consumed by the simulation process in the early stages of the design and the simplification of the predictions [19]. Several studies in the architectural field have been conducted typically integrating Artificial Neural Network (ANN) techniques for the prediction of building performance and environmental analysis [19] [9] [4] [5] [25]. In this research, we present a workflow where ML was trained and validated to predict incident solar radiation on an office building envelope using two algorithms: ANN [14], and Decision Tree (DT) [13]. The goal of conducting the simulation is to create a trained surrogate model to accurately predict solar radiation in significantly less time than would otherwise be required. Such an approach is worthwhile only if the number of cases that require prediction significantly outnumbers the number of cases needed for training.

2 RELATED WORK

In recent years, several studies have been conducted integrating the ML approach for predicting building performance, which includes building energy performance, estimating heating and cooling loads, daylighting, and solar predictions. Zhao and Magoulès (2012) agreed that the ML approach has proven to be efficient in the prediction of building performance [46]. Unlike conventional modelling methods, supervised ML has major benefits in terms of requiring less computation time and less effort and of being computationally less expensive [16]. Additionally, the accuracy and simplification of predictions has attracted researchers to investigate this possible alternative method for predicting building performance and occupant behaviour

SimAUD 2021 April 15-17 Virtually Hosted?

[42], and to replace building performance simulations by using data analytics [12].

A study by Paterson et al. (2013) created a design tool where ANN is integrated to predict energy consumption in real time [30]. The study focuses on school building design in England as a case study, and ANN was trained for prediction of the energy consumption of schools using the existing heating and electrical energy data of building stock to train the model. A more recent study by Asl et al. (2017) proposed a model called the Energy Model Machine (EMM) using ML algorithms, specifically ANN, to predict instant energy performance in the early stages of the design process [5]. The authors tested the EMM model in a medium-sized office building as a case study to demonstrate the usefulness of this method. The model generated 7,000 building design options with their energy performance, which helps designers make informed decisions during the conceptual design process. The researchers found that the use of ML to estimate energy performance during the process of design exploration and optimization is a feasible approach for achieving high-performance buildings.

A few studies have implemented neural networks (NN) to predict daylighting and illuminance. A study by Lee and Boubekri proposed a new method based on their exploration of the relationship between existing daylighting metrics and building design attributes [24]. Another study by Lopez and Gueymard (2007) used the NN approach for the prediction of luminous efficacy under cloudless conditions, which indicated the possibility of predicting the illuminances on surfaces based on solar irradiance measurements [26]. Additionally, in a study by Kazanasmaz et al. (2009), the authors applied NN-based modelling successfully to predict the horizontal illuminance in office buildings [19]. The study resulted in a low average error of 3% once it was compared to measured illuminances. A more recent study by Lorenz and Jabi (2017) analysed the efficiency of integrating supervised ML through using ANN to predict daylight autonomy levels for a typical office room [10]. The study found that more accurate results can be achieved when a large set of data is sufficiently trained.

3 MACHINE LEARNING (ML)

Machine learning (ML) is a branch of artificial intelligence (AI) that allows the software to learn without being explicitly programmed [8]. Mitchell (1997) defined it as a system that learns from past experience (i.e., data) to predict future performance [38]. In other words, ML could use existing data to predict or to respond to future data [15]. After training and learning, it is expected that the system should obtain a better predictive performance on the same trained task or related ones. In addition to the idea of self-improving automatically, ML also offers other advantages, such as collecting and clustering useful information from a large and complex set of data [23].

Recently, in the architecture field, ML has been proposed in several studies to estimate heating and cooling loads, building performance, energy consumption predictions, and architectural image recognition [34]. Nevertheless, the architecture field is considered one of the slowest industries to integrate ML, and it has resisted adopting it compared to other fields [21]. Carpo (2017) argued that architecture seems to be disregarding the potential of ML and its ability to predict performance, categorise large sets of data, and form optimization and advanced form findings [7].

4 ARTIFICIAL NEURAL NETWORK (ANN)

The artificial neural network (ANN) model has been widely utilized as a predictive tool in many fields [9]. The model was introduced by Mclloach-Hopfield in the early 1960s, but it started to develop more fully after 1985 [35]. The ANN model has the ability to deal with complex systems and nonlinear problems and is loosely inspired by the human brain [14]. The classic ANN model comprises units called neuros and is constructed in three parts, namely, the input layer, the output layer, and one or more hidden layers in between, and each of these hidden layers is composed of several neurons. Several researchers have applied the ANN method to predict and evaluate energy in buildings [43]. Some of these studies collected the dataset based on synthetic data simulation or based on real data. Keshtkarbanaeemoghadam et al. (2018) used a simulation-based approach to develop an ANN model, trained by a back-propagation algorithm (BP), for estimating the total heating energy demand of a shelter located in Iran. The study obtained the data by conducting 328 computer simulations using a Grasshopper plugin linked to the EnergyPlus engine. Different ANN models were examined with one or two hidden layers to select the most suitable architecture network. According to the results, the best ANN model had an MSE of 0.73, which indicates that the ANN model is a promising approach and can substitute other methods to predict the heating energy demand in buildings [20].

5 DECISION TREE (DT)

Decision tree (DT) is a method that has been commonly used for classification and prediction in many fields [41]. DT "uses a flowchart-like tree structure to segregate a set of data into various predefined classes, thereby providing the description, categorization, and generalization of given datasets" [44]. The DT model has advantages over other models because of its ease of use and the ability to predict accurately without requiring much computation time. While this method has the ability to process both numerical and categorical data, DT usually performs better with categorical than with numerical data [44]. A few applications have implemented DT techniques in relation to building studies [3]. Tso and Yau (2006) presented a comparison study between three modelling techniques to estimate average weekly electricity energy consumption in Hong Kong [40]. They found that both DT and ANN are applicable models compared to a regression model because of their understanding of energy consumption patterns and the prediction of energy usage. In another study by Haghighat et al. (2010), the researchers developed a predictive model to improve building energy performance based on the use of DT [44]. They applied the use of DT on a residential building to predict the energy use intensity (EUI) level. They concluded that the use of the DT method makes it possible to classify and predict the energy usage of the building accurately, which would lead to a high energy performance building.

6. METHODOLOGY

The methodological framework of this research is divided into two main steps: (1) simulation, and (2) machine learning (ML) prediction (Figure 1). In the first step, a generative parametric office tower and its urban context were designed and simulated using Ladybug, which is a plug-in tool for the Grasshopper software [32]. Moreover, in this step, the model settings were defined; these include setting up the parameters of the office tower, determining the design variables, and establishing the objective of the simulation, which is the solar radiation output. The ultimate goal of this phase was to create a large synthetic database of hourly solar radiation (KWh/m^2) data for training the model. The study selected the synthetic database approach because real data was not available. In the second step, NN and DT models were developed and evaluated for predicting the hourly solar radiation for a single closed office space within the office tower.

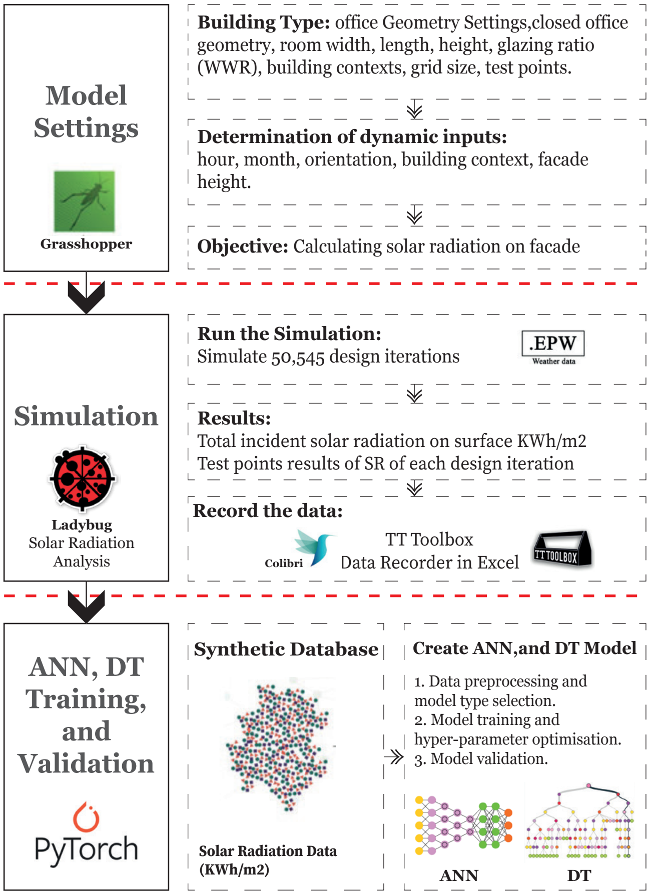

Model Settings
Grasshopper
Building Type: office Geometry Settings,closed office geometry, room width, length, height, glazing ratio (WWR), building contexts, grid size, test points.
Determination of dynamic inputs: hour, month, orientation, building context, facade height.
Objective: Calculating solar radiation on facade
Simulation
Ladybug
Solar Radiation Analysis
Run the Simulation: Simulate 50,545 design iterations
.EPW
Weather data
Results: Total incident solar radiation on surface KWh/m2
Test points results of SR of each design iteration
Record the data:
Colibri
TT Toolbox
Data Recorder in Excel
TT TOOLBOX
ANN, DT Training, and Validation
PyTorch
Synthetic Database
Solar Radiation Data (KWh/m2)
Create ANN,and DT Model
1. Data preprocessing and model type selection.
2. Model training and hyper-parameter optimisation.
3. Model validation.
ANN
DT

Figure 1. Framework of the study.

6.1 Establishing the Hypothetical Building

In this research, a generic mid-rise office building was developed as a case study in a hypothetical urban context located in the centre of Riyadh, Saudi Arabia. The office building has a height of 104 m, which is the median height found in the centre of the King Abdullah Financial District. This office building was located in a theoretical site and was assumed to be surrounded by several mid-rise office buildings, which created direct, diffuse, and reflected solar gains on the building surface (Figure 2) [39]. These solar gains would affect the annual energy demand of each office room of the building. The model was tested examining a hot-arid climate such as Riyadh city, where overheating is a crucial factor. The analysed office building has twenty floor levels with a fixed floor height of 4 m; this is representative of the common heights of offices found in the region. The dimensions of the layout and core area are fixed in all floors of the building as follows: (23.2 m * 23.2 m), with a total area of 538.20 m^2, the area of core services is 125.40 m^2 (11.20 m * 11.20 m), and the total gross area (GIA) of the office zones is 412.80 m^2.

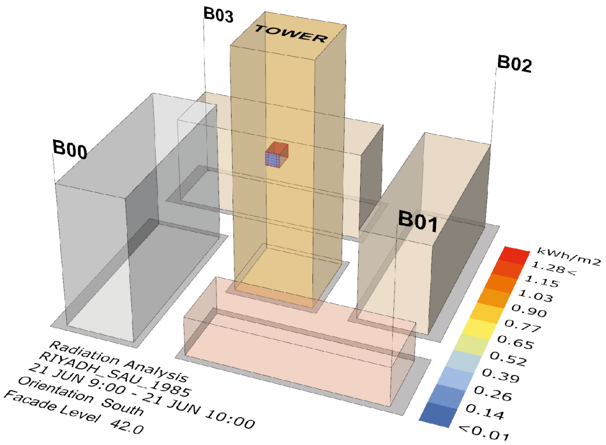


Figure 2. The 3D parametric urban context, that vary in each simulation.

6.2 Establishing the Typical Office

The research examined only a typical side-lit office room facing the main orientations (north, south, east, and west) with different floor levels that varied based on the surrounding contexts (Figure 3). This closed office room was designed with a fully glazed working environment, giving a window-wall ratio (WWR) of 80%. The spatial dimensions of the office room are 4 m wide by 6 m deep, making a rectangular zone with a floor to ceiling height of 4 m. The model incorporated most of the design conditions of a mid-rise office building.

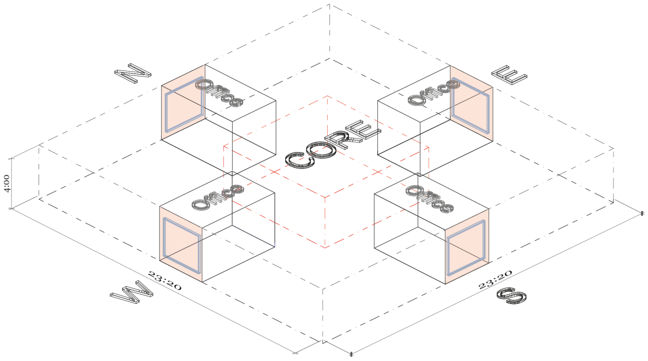


Figure 3. A single closed office space within the office tower facing main orientations.

6.3 Setting Up Solar Radiation Simulation

Computational tools can be applied effectively to gather quantitative information about both the building's performance and its design in the schematic design stage. Sadeghipour and Pak (2013) developed Ladybug and Honeybee, which are plug-in simulation tools for the Grasshopper platform [32]. These tools perform hourly calculations of different analyses, such as the total energy demand. In addition, Ladybug provides solar radiation analysis for calculating the energy collected on the building surface [22]. These plug-in tools are linked to different simulation engines, such as Daysim, Radiance, and EnergyPlus. In this experiment, simulation was conducted using the data from the weather file of Riyadh city, which were imported via the Ladybug plug-in tool [32]. The solar radiation that strikes the window surface of each office room was calculated considering the urban context variation. Since solar radiation differs based on different parameters, such as surrounding context, orientation, hour of the day, month, etc., a generative design process was conducted parametrically with varied parameters to simulate most of the design settings.

To that end, solar radiation analysis was performed considering the following main parameters: (1) Office operational time, which was considered to be from 8:00 am to 6:00 pm, (2) Day (21st of each month), which was constant in all the simulations, and (3) Month, which was selected seasonally (March, June, September, and December) throughout all the simulations [37]. (4) The building context varied in each simulation (low, medium, and high) to test solar radiation on all the levels and in all the main orientations (north, south, east, and west). The variation of heights of the surrounding contexts acted as one of the main features of geometrical variation in the study. In addition, the average height of the surrounding buildings was used parametrically to control the vertical location of the office room in each orientation in accordance with a lower than average, average, and higher than average height setting. This was meant to simulate the varying amounts of sunlight and daylight that the offices in a building receive. The vertical location of the office is calculated using the following formula.

$$a = \sum \mathrm{B00} + \mathrm{B01} + \mathrm{B02} + \mathrm{B03}/n$$

$$l = (a) * 0.50$$

$$h = (a) * 1.50$$

Where a= Average, l= Lower than average, h= Higher than average

B= Building Context, n= number of variables

In total, 324 different urban configurations were generated varying in height (low=12 m, medium=28 m, high=44 m). These are multiplied with 4 orientations, 13-day time hours, and 4 months. Figure (4) lists in detail these dynamic input parameters together with the fixed inputs used in this study to calculate the solar radiation collected on the building envelope.

The Colibri plug-in tool in Grasshopper was applied within the simulation workflow to step through all design variations automatically to create the dataset (Figure 5). Then, Colibri stored the result of the solar radiation data and its coordinates in an Excel spreadsheet [28].

Fixed Input Parameters

Input	Assigned Values
Site Location	Riyadh, Saudi Arabia
Space type	Shared Office Room
Room width	4.00 m
Room floor height	4.00 m
Room length	6.00 m
Glazing ratio	80 %
Grid size	0.80 * 0.80
Distance from base	0.01
Day	21st
Test Points coordinates	25 coordinates of X, Y, Z

Dynamic Input Parameters

Input	Assigned Values	No. of Iterations
Month	0,1,2,3	4
Hour	6:00 to 18:00	13
B00	L(12 m), M(28 m), H(44 m)	3
B01	L(12 m), M(28 m), H(44 m)	3
B02	L(12 m), M(28 m), H(44 m)	3
B03	L(12 m), M(28 m), H(44 m)	3
Orientation	South, West, North, East	4
Façade Height	l(0.50), a(1.00),h(1.50)	3
Total No of iterations		50,544

0= March, 1=June, 2=September, 3= December
B= Building Context Height , A total of 4 surrounding buildings
L=Low(12m), M=Medium(28m), H=Hight(44m)
l=lower than average, a=Average, h=Higher than average

Figure 4. Fixed and dynamic model parameters.

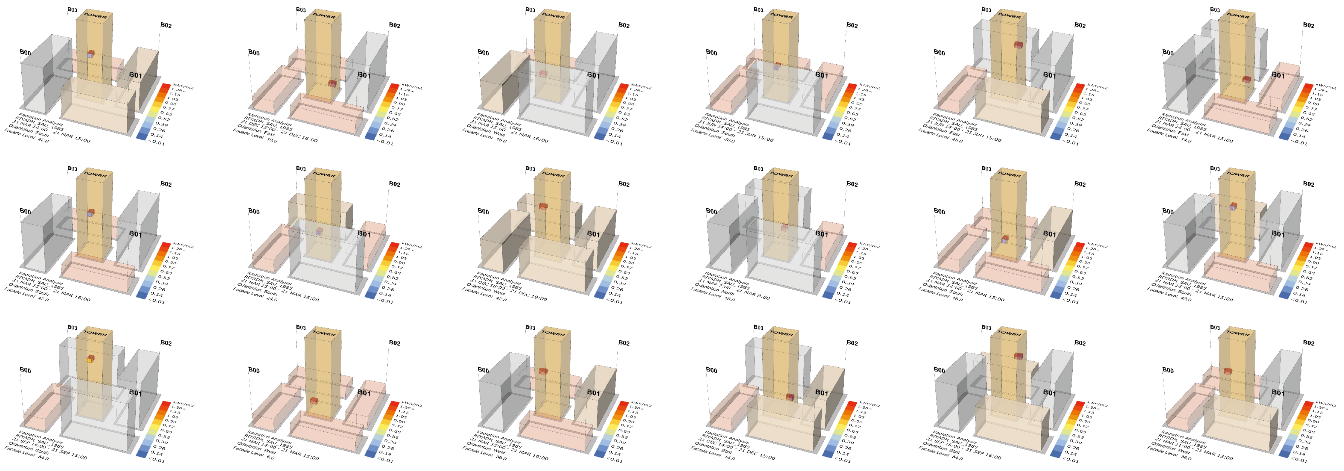


Figure 5. Sample of the solar radiation data generated automatically in the simulation process.

As stated in the literature, employing sufficient data is essential to achieving a high accurate predictive model that can predict the hourly solar radiation of the surface of the office room [17]. For this purpose, a total of 50,545 solar radiation iterations were generated in the simulation process examining different orientations. Figure 6 shows an example of some samples of the conducted simulation for solar radiation results in two different orientations, and two different hours of the day.

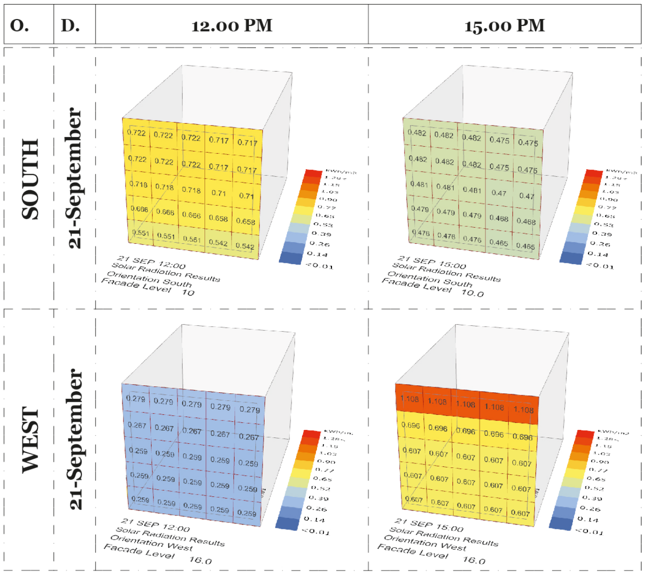

O.
D.
12.00 PM
15.00 PM
SOUTH
21-September
WEST
21-September
21 SEP 12:00 Solar Radiation Results Orientation South Facade Level 10
21 SEP 15:00 Solar Radiation Results Orientation South Facade Level 10.0
21 SEP 12:00 Solar Radiation Results Orientation West Facade Level 16.0
21 SEP 15:00 Solar Radiation Results Orientation West Facade Level 16.0

Figure 6. An example of two different cases of solar radiation.

The output of each iteration consists of a number of test points that fall on the tested surface with X, Y, and Z coordinates to perform the radiation analysis, and these values are measured in KWh/m^2 (Figure 7). The total number of test points is 1,263,600 (25 test points of each surface * 50,545 total number of iterations). The total radiation results in KWh/m^2 are calculated through the mass addition of results at each of the test points multiplied by the area of the face that the test point is representing.

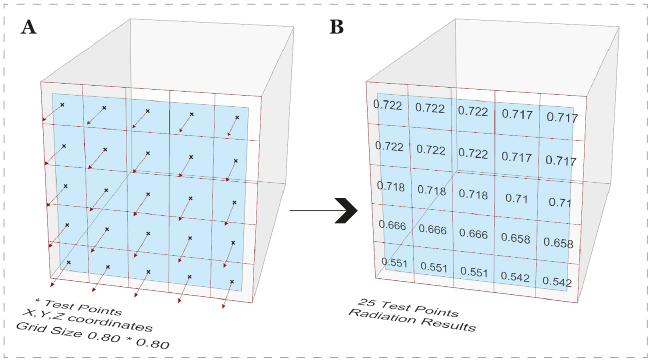


Figure 7. (A) Test points coordinates**,** (B) Radiation results based on each test point.

7 MODEL TRAINING AND TESTING

In this section, we discuss the data used for the modelling and the machine learning (ML) algorithms used.

7.1 Generated Data

The data generated in the simulation stage were imported into two ML algorithms (ANN) and (DT) to create a model for training and testing the simulation results. A total of twelve variables were used as input parameters to train the two models. The output of the data is the solar radiation of the corresponding coordinates.

7.2 Data Pre-Processing

Among the input data features used for the modelling, Hour, Month, Building contexts (B00, B01, B02, B03), Façade floor level, Orientation, and Façade level height are considered categorical features, and they are one hot encoded which is a common way of converting categorical inputs into a suitable format for ML models [33]. The remaining input features are x/y/z coordinates of the test points. They are treated as continuous inputs and pre-processing is not applied to them. The output is the solar radiation of the corresponding location. The input data feature, and ranges of each input are illustrated in Table (1).

Input	Input Neuron Type	Data Range
Hour	Discrete	6 to 18 in steps of 1
Month	Discrete	0,1,2,3
B00	Discrete	0,1,2
B01	Discrete	0,1,2
B02	Discrete	0,1,2
B03	Discrete	0,1,2
Facade Floor Level	Discrete	0,1,2
Orientation	Discrete	0,1,2,3
Façade Height	Discrete	6 to 66
x-coordinate	Continuous	[-11.08, 12.28]
y-coordinate	Continuous	[-13.06, 10.26]
z-coordinate	Continuous	[6.40, 69.60]

Table 1: The input data used for the machine learning modeling.

7.3 Neural Network Modeling

Artificial Neural Network (ANN) are used for the modeling problem in the form of a regression learning. The performance of the network is evaluated with the root mean square error (RMSE), mean absolute error (MAE) and R2-score (R2) which are calculated using the following formulae. $\mathrm{RMSE} = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\ (y_i - f(x_i))^2}$,

$$\mathrm{MAE} = \frac{1}{|n|}\sum_{i=1}^{|n|} |y(i) - f(x_i)|\,,\ R^2 = 1 - \frac{\sum_{i=1}^{|n|}(y(i) - f(x_i))^2}{\sum_{i=1}^{|n|}(y(i) - \bar{y})^2}$$

In the above equations, we assume there are *n* number of testing data points, y_i is the output of the i-th data point corresponding to the input x_i, *f*(x_i) is the predicted value of the i-th data point where *f(x)* is the function approximated by the neural network and $\bar{y} = \frac{1}{|n|}\sum_{i=1}^{|n|} f(x_i)$. In an ideal modeling case, we expect the RMSE and MAE values to be zero and the R^2score to be 1. The one hot encoded categorical features are fed into an embedding layer. The purpose of this layer is to give a vector embedding to the one hot encoded input rather than using them as such. This approach is found to give more representational power for the categorical inputs. Similarly, the continuous features are fed into the batch normalizing layer [18]. Its purpose is to make the continuous data follow the same probability distribution so that the learning of the network is optimized. For the non-linear activation of the inputs in the neurons, the RectifiedLinear Unit (ReLU) function defined as *f(x) = max(0, x)* is used [27]. The output of the ReLU is batch normalized, and then the dropout regularization is applied. Dropout is a mechanism to ensure the generalization capability of the network by avoiding overfitting [36]. For each layer of the network, the data are processed as – linear layer à ReLu activation à batch normalization à dropout. The output neuron of the network is a trivial neuron where no activations are applied, and the output is taken from the previous layer (Figure 8). Implementation of the network was done using the Pytorch framework [29]. The dropout rate used for the experiment was 0.2.

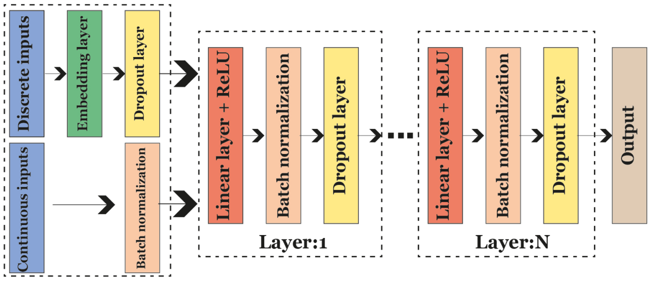


Figure 8. ANN Components.

7.4 Random Forest Modeling

Random forest (RF) for regression is used for modelling the data. The inputs to the RF model are the one hot encoded categorical features and x/y/z coordinates of a location. The output is the solar radiation of the corresponding location. The performance of the network is evaluated with the RMSE, MAE, and R2-score.

8 OPTIMIZATION OF ANN - RESULTS

For the ANN modeling, a suitable architecture has to be selected. Along with this, there are hyper-parameters such as learning rate, batch-size etc that require fine-tuning. For choosing the right architecture for ANN an experiment is designed based on k-fold cross validation; the description is given below.

8.1 K-fold Cross Validation

The purpose of the k-fold cross validation is to choose the right architecture for the ANN in terms of the number of layers required and the number of neurons in each layer. The architectures used for the experiment are 1 ,2, 3 and 4 hidden layer networks. Each network was tested with 64, 128, 256, 512, and 1024 neurons in each layer. For this experiment, the data is split into training, validation and test sets. The models are tested using the training and validation data splits and the model finally chosen is tested with the test data. It has to be noted that the test data is treated as unseen data and it was not used for selecting the right architecture for ANN. The details of the cross-validation experiment are given below.

8.1.1 Data Split

Initially, the whole dataset is split into training, validation, and testing sets: 80% of the data is assigned to the training set, 6.67% to the validation set, and the remaining 13.37% to the testing set. The k-fold cross validation is then conducted on the training fold. The value of k chosen is 5. Note that while doing k-fold cross validation, one among the fold becomes the testing set and the remaining become the training set. In this case, one-third of the testing case will be reserved as a validation set for that particular instance of the validationprocedure. For cross-validation experiments, the learning rate is fixed at 0.01, the dropout rate at 0.2, and the batch size at 16,000. The batch size has taken this value as it optimizes the hardware utilization. The experiments are run for 100 epochs with an early stopping criterion of 10 epochs.

8.1.2 Results

As we increased the number of layers, the RMSE decreased. This was expected since the representation capabilities of the NN were also increasing as we increased the number of layers. However, there was a slight increase of RMSE for 4-layer networks. A similar trend could be observed with the MAE and R2 scores as well. The architecture selected after the 5-fold cross validation experiment is a 3-layer network with 256 neurons in each layer. The result of the k-fold cross validation is given in Table 2.

No: of layers	No: of neurons	RMSE	MAE	R2
1	512	0.01842	0.07976	0.7334
2	256	0.01524	0.06716	0.7794
3	256	0.01509	0.06596	0.7832
4	128	0.01539	0.06675	0.7789

Table 2: Summarized architecture optimization of ANN.

8.2 Testing the architecture

The 3-layer network with 256 neurons in each layer which was selected through k-fold cross validation was applied with the validation set and test set. For this purpose, a new model was built using the entire training set. This model was tested with the test set. The results obtained are as follows: RMSE=0.011415, MAE=0.052188, and R^2-score=0.831315. For the experiment, the learning rate used was 0.02, the dropout rate was 0.2, the number of epochs was 100, and the batch size was 16,000. Figure 9 shows a sample prediction heatmap made by ANN for a set of 25 coordinates to test the model.

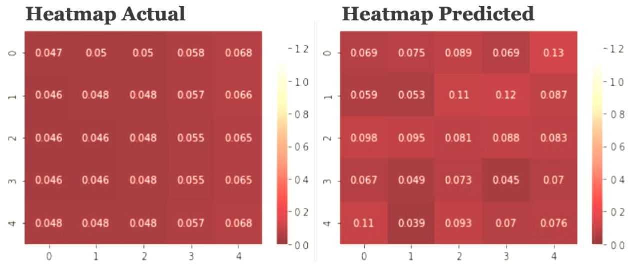


Figure 9. The actual and predicted solar radiation values by ANN.

9 OPTIMIZATION OF RANDOM FOREST - RESULTS

For random forest modelling there are several hyper-parameters involved as follows: (1) *Number of trees*: The number of trees indicates the number of individual decision tree estimators. Its value is tuned from the set {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. (2) *Bootstrap:* Bootstrap is a process of random sampling from the training data with replacement. This procedure helps reduce the high variance of the random forest models and prevents them from over-fitting [6]. If the bootstrap option is not enabled, the random forest is learned using the whole data. (3) *Minimal Cost-Complexity Pruning parameter:* Tree pruning is a procedure to avoid over-fitting in random forest models [11]. An experiment was done by varying the pruning hyper-parameter ccp-alpha in the set [0, 0.001, 0.002, 0.003, 0.004, 0.005].

9.1 K-fold Cross Validation

To find the proper values for the hyper-parameters, an experiment was conducted similar to the one described in Section 8. The dataset was split into 80% for training and 20% for testing. Then, 5-fold cross validation was applied on the training set alone to fix the hyper-parameters. To test the model in the testing set, a model was built using the entire training set whose hyper-parameters were those fixed by the k-fold cross validation.

9. 2 Results

The results of the k-fold cross validation are given in Table 3. The results in the table are given in a summarized form, that is, only the best result obtained for each of the options of number of trees is given.

Sl No	Trees	RMSE	MAE	R^2-score
1	10	3.52E-04	4.42E-03	9.92E-01
2	20	3.48E-04	4.41E-03	9.93E-01
3	30	3.48E-04	4.40E-03	9.95E-01
4	40	3.48E-04	4.39E-03	9.95E-01
5	50	3.45E-04	4.40E-03	9.95E-01
5	60	3.52E-04	4.48E-03	9.94E-01
7	70	3.45E-04	4.39E-03	9.92E-01
8	80	3.54E-04	4.57E-03	9.93E-01
9	90	3.41E-04	4.38E-03	9.95E-01
10	100	3.50E-04	4.40E-03	9.95E-01

Table 3: The k-fold cross validation results for DT

From the results, there is no significant variation in the RMSE value as the number of trees increases. In addition, the observation is true for other metrics. The best performance is observed when the ccp-alpha value is 0.0 irrespective of the number of trees and the Bootstrap option. When a non-zero value is specified, there is a greater drop in the performance. This is very evident, as the R^2-score is 0. The performance when the bootstrap option is enabled is better than with the models where it is disabled. The best result is observed when the number of trees is 80, the ccp-alpha value is 0, and the Bootstrap option is enabled. In this setting, the RMSE value is 0.000354, the MAE is 0.00457, and the R^2-score is 0.993. The final test results are as follows. The RMSE is 0.000514, the MAE is 0.00661, and the R^2-score is 0.99228. Figure 10 shows a sample prediction heatmap made by DT for a set of 25 coordinates to test the model.

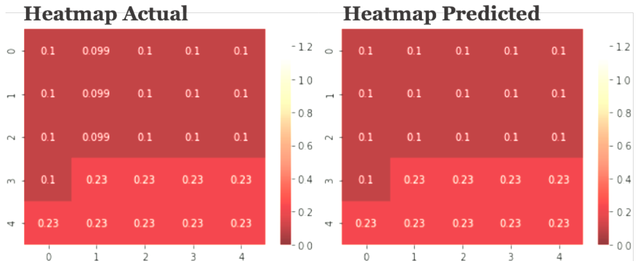


Figure 10. The actual and predicted solar radiation values by DT.

10 DECISION TREE SURROGATE MODEL TESTING WITH NEW SCENARIO

After the training and validation had been conducted with ANN and DT, the trained DT surrogate model was imported into the Grasshopper interface using GH CPython to predict solar radiation for a new scenario. GH CPython was developed by Mahmoud M. Abdelrahman for the integration of the Grasshopper and CPython languages [1]. This plug-in tool allows users to import Python libraries such as NumPy, SciPy, Matplotlib, pandas, Scikit-learn, Pytorch, etc. into Grasshopper and link the trained model to predict for a new scenario. We created a set of new test scenarios that is not part of the data used to build machine learning models. The new design considers several office towers with new urban contexts as shown in Figure 11. The output results of solar radiation with the DT surrogate model are very close to the simulation prediction results shown in Figure 12. To that end, our approach will be a cost-effective solution for the cases that require large scale generation of simulation data. The implementation codes and files are available in this link: https://github.com/archammar/Solar-radiation-prediction-for-office-tower-using-machine-learning

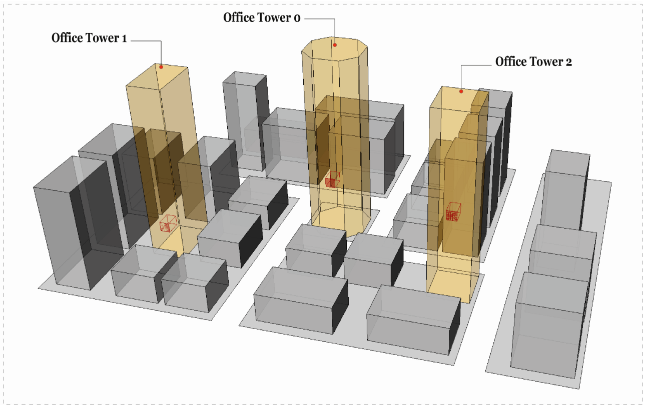
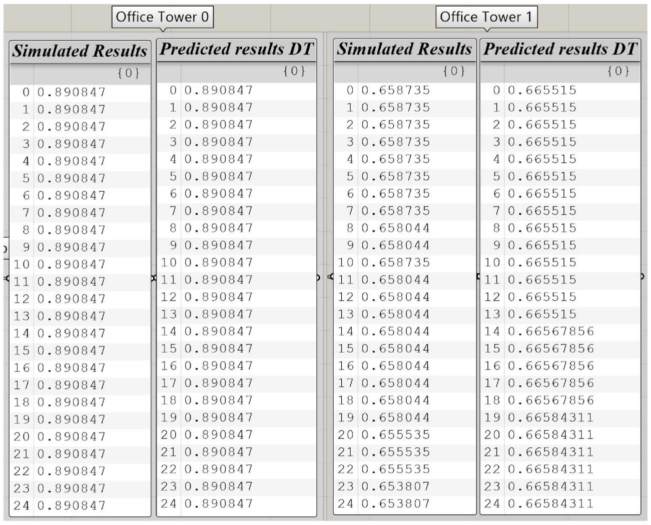


Figure 11. New scenario with several office towers.



Office Tower 0		Office Tower 1	
Simulated Results	*Predicted results DT*	*Simulated Results*	*Predicted results DT*
{0}	{0}	{0}	{0}
0 0.890847	0 0.890847	0 0.658735	0 0.665515
1 0.890847	1 0.890847	1 0.658735	1 0.665515
2 0.890847	2 0.890847	2 0.658735	2 0.665515
3 0.890847	3 0.890847	3 0.658735	3 0.665515
4 0.890847	4 0.890847	4 0.658735	4 0.665515
5 0.890847	5 0.890847	5 0.658735	5 0.665515
6 0.890847	6 0.890847	6 0.658735	6 0.665515
7 0.890847	7 0.890847	7 0.658735	7 0.665515
8 0.890847	8 0.890847	8 0.658044	8 0.665515
9 0.890847	9 0.890847	9 0.658044	9 0.665515
10 0.890847	10 0.890847	10 0.658735	10 0.665515
11 0.890847	11 0.890847	11 0.658044	11 0.665515
12 0.890847	12 0.890847	12 0.658044	12 0.665515
13 0.890847	13 0.890847	13 0.658044	13 0.665515
14 0.890847	14 0.890847	14 0.658044	14 0.66567856
15 0.890847	15 0.890847	15 0.658044	15 0.66567856
16 0.890847	16 0.890847	16 0.658044	16 0.66567856
17 0.890847	17 0.890847	17 0.658044	17 0.66567856
18 0.890847	18 0.890847	18 0.658044	18 0.66567856
19 0.890847	19 0.890847	19 0.658044	19 0.66584311
20 0.890847	20 0.890847	20 0.655535	20 0.66584311
21 0.890847	21 0.890847	21 0.655535	21 0.66584311
22 0.890847	22 0.890847	22 0.655535	22 0.66584311
23 0.890847	23 0.890847	23 0.653807	23 0.66584311
24 0.890847	24 0.890847	24 0.653807	24 0.66584311

Figure 12. Comparison between simulated solar radiation results using Ladybug and predicted results using DT surrogate model.

11 CONCLUSION AND RECOMMENDATIONS

This paper aimed to find an alternative approach for solar analysis studies using a machine-learning (ML) method. This technique is beneficial for saving time that is mostly consumed during the simulation process to inform the design decision of related building envelope studies, such as the design of the shading system, glazing ratio, and PV envelope solar systems, etc. Initially, we created a large synthetic data through model - simulation approach of solar radiation analysis. Then, we fed these datasets into two ML algorithms to train and test the model. To accomplish the best prediction result, sensitivity analysis tests were performed using different parameters.

After the experiment with these different parameters, the best result achieved for the neural network in terms of RMSE is 0.011415 and for the random forest is 0.000514. While both methods provided an acceptable level of accuracy, the performance by DT is significantly higher than that of ANN as given in Table 4. This is due to the following reasons:

- Most of the inputs are categorical. This is a perfect learning scenario for the random forest. Since the decision trees are highly suitable for categorical inputs in the learning setting, random forest modelling has a significant advantage over neural network modelling.
- A random forest is an ensemble algorithm. Ensemble algorithms are a class of ML algorithms where a set of learning programs are combined together to give accurate predictions [11]. Hence its performance was significantly better than that of neural networks.

The generalizability of either DT or NN to a new modelling problem is dependent on the nature of the inputs and the modelling tasks. However, based on our studies, we observe that if the input data is mostly categorical, the DT algorithm could perform better than ANN.

Performance metric	Neural Network	Decision Tree
RMSE	0.011415	0.000514
MAE	0.052188	0.00661
R2-score	0.831315	0.99228

Table 4: Performance comparison of NN and DT.

One of the main limitations of this study is the unavailability of real-life data, so a simulation was performed to produce data. Additionally, the case study focused only on a hot climate region and tall office towers within an urban context, so the findings of this study cannot be generalised to other climates. Although we were able to predict solar radiation on a vertical façade based on the variation in the urban context, façade level heights, orientations, and different hours, further research is needed to examine other parameters, such as the materials of the urban context to predict a more accurate result.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Asif Salim, IIST for his help with the pytorch programming environment.

REFERENCES

[1] Abdelrahman, M.M. 2017. Enhancing Computational Design with Python high performance scientific libraries : Integration of Grasshopper and CPython language. November (2017), 2–3. DOI:https://doi.org/10.13140/RG.2.2.27230.33600/1.

[2] Ahmad, M.H. et al. 2005. Impact Of Solar Radiation On High-Rise Built Form In Tropical Climate. *NSEB – National Seminar on Energy in Buildings, UiTM, Shah Alam– National Seminar on Energy in Buildings, UiTM, Shah Alam*. (2005), 1–9.

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

[4] Ascione, F. et al. 2017. Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy*. 118, (2017), 999–1017. DOI:https://doi.org/10.1016/j.energy.2016.10.126.

[5] Asl, M.R. et al. 2017. Energy Model Machine (EMM). 2, (2017), 277–286.

[6] Breiman, L. 2019. Random forests. *Random Forests*. (2019), 1–122. DOI:https://doi.org/10.1201/9780367816377-11.

[7] Carpo, M. 2017. *The second digital, turn design beyond intelligence*. The MIT Press.

[8] Chakraborty, D. and Elzarka, H. 2019. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation*. 12, 2 (2019), 193–207. DOI:https://doi.org/10.1080/19401493.2018.1498538.

[9] Chew, M.Y.L. et al. 2004. A neural network approach to assessing building façade maintability in thr tropics. *Construction Management and Economics*. 22, 6 (2004), 581–594. DOI:https://doi.org/10.1080/01446190310001631019.

[10] Clara Lorenz and Jabi, W. 2017. Interna onal Conference for Sustainable Design of the Built Environment SDBE 2017. December (2017).

[11] Dietterich, T.G. 2000. Ensemble Methods in Machine Learning. *Multiple Classifier Systems* (Berlin, Heidelberg, 2000), 1–15.

[12] Geyer, P. and Singaravel, S. 2018. Component-based machine learning for performance prediction in building design. *Applied Energy*. 228, October 2017 (2018), 1439–1453. DOI:https://doi.org/10.1016/j.apenergy.2018.07.011.

[13] Han, J. et al. 2014. *Data mining: Data mining concepts and techniques*.

[14] Haykin, S. 1992. *Neural networks and learning*.

[15] Hopmann, C. et al. 2017. *Machine Learning to improve indoor climate and building energy performance Energy Engineering and Management Examination Committee*.

[16] Huang, H. et al. 2015. A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings. *Energy and Buildings*. 97, (2015), 86–97. DOI:https://doi.org/10.1016/j.enbuild.2015.03.045.

[17] Ilbeigi, M. et al. 2020. Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm. *Sustainable Cities and Society*. 61, June (2020), 102325. DOI:https://doi.org/10.1016/j.scs.2020.102325.

[18] Ioffe, S. and Szegedy, C. 2016. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Journalism Practice*. 10, 6 (2016), 730–743. DOI:https://doi.org/10.1080/17512786.2015.1058180.

[19] Kazanasmaz, T. et al. 2009. Artificial neural networks to predict daylight illuminance in office buildings. *Building and Environment*. 44, 8 (2009), 1751–1757. DOI:https://doi.org/10.1016/j.buildenv.2008.11.012.

[20] Keshtkarbanaeemoghadam, A. et al. 2018. Estimation and optimization of heating energy demand of a mountain shelter by soft computing techniques. *Sustainable Cities and Society*. 41, June (2018), 728–748. DOI:https://doi.org/10.1016/j.scs.2018.06.008.

[21] Khean, N. et al. THE INTROSPECTION OF DEEP NEURAL NETWORKS -. 2, Ml, 237–246.

[22] Kim, D. et al. 2012. A Symbiotic Interaction of Virtual and Physical Models in Designing Smart Building Envelope. 2, (2012), 633–642.

[23] Kwok, J.T. et al. 2015. Machine Learning. *Springer Handbook of Computational Intelligence*. J. Kacprzyk and W. Pedrycz, eds. Springer Berlin Heidelberg. 495–522.

[24] Lee, J. et al. 2019. Impact of building design parameters on daylighting metrics using an analysis, prediction, and optimization approach based on statistical learning technique. *Sustainability (Switzerland)*. 11, 5 (2019). DOI:https://doi.org/10.3390/su11051474.

[25] Lee, S. et al. 2019. Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. *Energies*. 12, 4 (2019). DOI:https://doi.org/10.3390/en12040608.

[26] López, G. and Gueymard, C.A. 2007. Clear-sky solar luminous efficacy determination using artificial neural networks. *Solar Energy*. 81, 7 (2007), 929–939. DOI:https://doi.org/10.1016/j.solener.2006.11.001.

[27] Nair, V. and Hinton, E., G. 2017. Rectified Linear Units Improve Restricted Boltzmann Machines. *Journal of Applied Biomechanics*. 33, 5 (2017), 384–387. DOI:https://doi.org/10.1123/jab.2016-0355.

[28] Natanian, J. et al. 2019. A parametric approach to optimizing urban form, energy balance and environmental quality: The case of Mediterranean districts. *Applied Energy*. (2019). DOI:https://doi.org/10.1016/j.apenergy.2019.113637.

[29] Paszke, A. et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS (2019).

[30] Paterson, G. et al. 2017. Energy use predictions with machine learning during architectural concept design. *Science and Technology for the Built Environment*. 23, 6

(2017), 1036–1048. DOI:https://doi.org/10.1080/23744731.2017.1319176.

[31] Purnama, M. and Sutanto, D. 2017. Dynamic facade module prototype development for solar radiation prevention in high rise building. *Iopscience.Iop.Org*. 8, February 2018 (2017), 68–74. DOI:https://doi.org/10.1088/1755-1315.

[32] Sadeghipour Roudsari, M. et al. 2013. Ladybug: a Parametric Environmental Plugin for Grasshopper To Help Designers Create an Environmentally-Conscious Design. *13th Conference of International building Performance Simulation Association*. (2013), 3129–3135.

[33] Seger, C. 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. *Degree Project Technology*. (2018), 41.

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

[35] Simpson, P.K. 1990. *Artificial neural systems: Foundation, Paradigms, Applications, and Implementations*. Pergamon, Elmsford, NY.

[36] Srivastava, N. et al. 2016. Dropout: A Simple Way to Prevent Neural Networks from Overfittin. *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering, ICAEES 2016*. 15, (2016), 520–525. DOI:https://doi.org/10.1109/ICAEES.2016.7888100.

[37] Tabadkani, A. et al. 2018. Daylighting and visual comfort of oriental sun responsive skins : A parametric analysis. (2018), 663–676.

[38] Thomas M. Mitchell 1997. *Machine Learning Book*. McGraw-Hill, Inc. New York, NY, USA ©1997.

[39] Trigaux, D. and De Troyer, F. 2015. A design tool to optimize solar gains and energy use in neighbourhoods. *PLEA 2015 book of proceedings*. September (2015).

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

[41] Tung, K.Y. et al. 2005. Mining the Generation Xers' job attitudes by artificial neural network and decision tree - Empirical evidence in Taiwan. *Expert Systems with Applications*. 29, 4 (2005), 783–794. DOI:https://doi.org/10.1016/j.eswa.2005.06.012.

[42] Wang, Z. and Srinivasan, R.S. 2017. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*. 75, October 2016 (2017), 796–808. DOI:https://doi.org/10.1016/j.rser.2016.10.079.

[43] Westermann, P. and Evins, R. 2019. Surrogate modelling for sustainable building design – A review. *Energy and Buildings*. 198, (2019), 170–186. DOI:https://doi.org/10.1016/j.enbuild.2019.05.057.

[44] Yu, Z. et al. 2010. A decision tree method for building energy demand modeling. *Energy and Buildings*. 42, 10 (2010), 1637–1646. DOI:https://doi.org/10.1016/j.enbuild.2010.04.006.

[45] Zell, E. et al. 2015. Assessment of solar radiation resources in Saudi Arabia. *Solar Energy*. 119, (2015), 422–438. DOI:https://doi.org/10.1016/j.solener.2015.06.031.

[46] Zhao, H.X. and Magoulès, F. 2012. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*. 16, 6 (2012), 3586–3592. DOI:https://doi.org/10.1016/j.rser.2012.02.049.