

The use of statistical and machine learning tools to accurately quantify the energy performance of residential buildings

Dina M. Ibrahim^{1,2}, Abdulbasit Almhafdy³, Amal A. Al-Shargabi¹, Manal Alghieth¹, Ahmed Elragi⁴ and Francisco Chiclana⁵

¹ Department of Information Technology, College of Computer, Qassim University, Buraydah, Qassim, Saudi Arabia

² Department of Computers and Control Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt

³ Department of Architecture, College of Architecture and Planning, Qassim University, Buraydah, Qassim, Saudi Arabia

⁴ Department of Civil Engineering, College of Engineering, Qassim University, Buraydah, Qassim, Saudi Arabia

⁵ Institute of Artificial Intelligence (IAI), Faculty of Technology, De Montfort University Leicester, Leicester, Leicester, United Kingdom

ABSTRACT

Prediction of building energy consumption is key to achieving energy efficiency and sustainability. Nowadays, the analysis or prediction of building energy consumption using building energy simulation tools facilitates the design and operation of energy-efficient buildings. The collection and generation of building data are essential components of machine learning models; however, there is still a lack of such data covering certain weather conditions. Such as those related to arid climate areas. This paper fills this identified gap with the creation of a new dataset for energy consumption of 3,840 records of typical residential buildings of the Saudi Arabia region of Qassim, and investigates the impact of residential buildings' eight input variables (Building Size, Floor Height, Glazing Area, Wall Area, window to wall ratio (WWR), Win Glazing U -value, Roof U -value, and External Wall U -value) on the heating load (HL) and cooling load (CL) output variables. A number of classical and non-parametric statistical tools are used to uncover the most strongly associated input variables with each one of the output variables. Then, the machine learning Multiple linear regression (MLR) and Multilayer perceptron (MLP) methods are used to estimate HL and CL, and their results compared using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and coefficient of determination (R^2) performance measures. The use of the IES simulation software on the new dataset concludes that MLP accurately estimates both HL and CL with low MAE, RMSE, and R^2 , which evidences the feasibility and accuracy of applying machine learning methods to estimate building energy consumption.

Submitted 14 October 2021
Accepted 29 December 2021
Published 26 January 2022

Corresponding author
Dina M. Ibrahim,
dina.mahmoud@f-eng.tanta.edu.eg

Academic editor
Imran Ashraf

Additional Information and
Declarations can be found on
page 26

DOI 10.7717/peerj-cs.856

© Copyright
2022 Ibrahim et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Artificial Intelligence, Data Science

Keywords Buildings characteristics, Cooling load, Heating load, Energy consumption, Statistical analysis

INTRODUCTION

Research on building energy consumption is motivated by the recently growing concerns on energy waste and its negative impact on the environment. When designing efficient buildings, it is essential to calculate their cooling load (CL) and heating load (HL) in order to specify the required cooling and heating equipment to achieve comfortable indoor air conditions. Architects and building designers require information about building characteristics, conditioned spaces (occupancy and activity level), climate, and intended usage (residential, industrial) to estimate the CL and HL of the building. Buildings have five distinct characteristics: environment, utilities, community, occupants, and building system (Wang et al., 2017). The environmental characteristics of a building are among the main aspects or conditions that can affect its energy consumption, *i.e.* contribute to sustainability and energy efficiency. Therefore, this study focuses on buildings characteristics such as wall envelope, window, and orientation.

In the literature, buildings' characteristics have been described as "variables" (Tsanas & Xifara, 2012), "forms" (Li et al., 2019), "components" (Geyer & Singaravel, 2018), "shapes and characteristics" (Ciulla et al., 2019), and "features" (Seyedzadeh et al., 2019). Physical and non-physical factors can be used to categorize the characteristics of buildings. A window to wall ratio, for example, is a physical element of a building that is related to size, while glazing properties (*e.g.* *U*-value) are an example of physical elements of a building that are related to materials. The orientation of a building, which is determined by the cardinal and intercardinal building directions, is an example of non-physical factors.

The building characteristics in related studies can be categorized into five groups: wall variables, glazing variables, roof variables, form variables, and orientation. Glazing variables are a major architectural elements that identify the building's features and they have a significant impact on energy performance (Tien Bui et al., 2019; Yeom et al., 2020). Five different building envelope parameters have been used to address glazing: area, area distribution, window to wall ratio (WWR), window to ground ratio (WGR), and *U*-value. Furthermore, when looking at each variable separately, orientation is the variable most investigated in AI research studies. Most buildings' energy prediction studies, such as Yeom et al. (2020), Moayedi et al. (2019), Navarro-Gonzalez & Villacampa (2019), Seyedzadeh et al. (2019) and Sadeghi et al. (2020), conducted their experiments on a dataset, created by Tsanas & Xifara (2012) of 768 records and eight characteristics (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and distribution) used as predictors to estimate the energy consumption of the buildings.

In addition, a small number of other studies have used larger datasets. To name some of them (Himeur et al., 2020a), reviewed and examined thirty-one existing datasets based on various features such as geographical locations and rate sampling. The authors proposed a novel dataset, namely, Qatar University dataset which can be useful for any future training or testing anomaly detection algorithms. Another future direction of applying the datasets in several utilizations such as machine learning was also proposed. In addition, Li et al. (2020) used 539, 42, and 153 datasets of residential buildings,

residential blocks and public buildings respectively. The authors highlighted the buildings key determinants that affect the urban building energy usage, e.g. orientation, height to canyon width perimeter-to-area ratio. (Xu & Chen, 2020), collected datasets of energy consumption from various houses in British Columbia, Canada, for 2 years. The aim was to detect anomaly energy performance in buildings. (Pham et al., 2020), used five datasets from five buildings of 1 year with an hourly resolution of energy consumption for evaluating ML-based energy prediction model. By utilizing the historical datasets, Random Forests showed good accuracy in energy prediction. (Himeur et al., 2020c), validated a recognition system based on a non-intrusive appliance model using resampled data recording in power consumption with 30,000 patterns length. The proposed model showed high accuracy in appliance recognition performance.

However, all the above mentioned studies were not constructed based on the building characteristics which emerges the gap in the existing buildings envelope based datasets. (Himeur et al., 2020b) has stated that the lack of real or well-validated datasets is one of the main obstacles that stand before anomaly prediction and detection of energy consumption in buildings. Highlighting energy output has gone through various investigations, and yet, there are still difficulties in identifying the energy performance pattern, abnormalities. Thus, this study creates a new dataset of 3,840 typical family houses in the Qassim region of Saudi Arabia, and corresponding eight characteristics to predict energy consumption, which is to be available online for public use.

Based on the created dataset, a number of classical and non-parametric statistical tools are first used to uncover the most strongly characteristics (input variables) with HL and CL (output variables). Then, two machine learning methods, the Multiple linear regression (MLR) and the Multilayer perceptron (MLP), are used to estimate HL and CL, and their results are compared using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and coefficient of determination (R^2) performance measures. The use of the IES<VE> simulation software on the new dataset concludes that MLP accurately estimates both HL and CL with low MAE, RMSE, and R^2 , which evidences the feasibility and accuracy of applying machine learning methods to estimate building energy consumption. Thus, the main contributions of this study are:

1. A new dataset of 3,840 arid climate residential buildings and corresponding eight characteristics to predict energy consumption is made publicly available.
2. *In silico* experiments on the developed dataset evidence feasibility and accuracy of applying machine learning methods to estimate building energy consumption.

The remaining of the paper is structured as follows. “Existing Datasets of Energy Consumption In Residential Buildings” presents an overview of the existing dataset used in the literature. “Methodology” details the methodology implemented to create and analyze the new dataset. “Methods and Statistical Analysis Results” reports on the results of the dataset analysis using both statistical methods and machine learning methods. “Results and Discussions” discusses the obtained results and finally “Concluding Remarks and Future Research Directions” concludes the paper.

Existing datasets of energy consumption in residential buildings

The application of machine learning on building energy prediction is extensively addressed in the literature ([Zhang et al., 2021](#)). However, most of these studies focus on the algorithm implemented, while the dataset used is often overlooked. In [Tsanas & Xifara \(2012\)](#), Tsanas and Xifara presented a dataset on eight building characteristics (input variables X1–X8): Surface Area, Overall Height, Roof Area, Relative Compactness, Wall Area, Distribution of Glazing Area, Orientation, Area of Glazing; as predictors of buildings' energy consumption target variables (Y1–Y2): Heating Load and Cooling Load.

Many researches have used the [Tsanas & Xifara \(2012\)](#) dataset for various energy prediction models in various regions using 12 different building shapes simulated in Autodesk Ecotect Analysis too (see [Table 1](#)). [Kumar, Pal & Singh \(2018\)](#) used the data for residential buildings in California, [Navarro-Gonzalez & Villacampa \(2019\)](#) in Alicante, Spain, and [Roy et al. \(2020\)](#) in Athens, Greece. [Ciulla et al. \(2019\)](#) employed nonresidential building simulation data from seven countries: Germany, Spain, the United Kingdom, Belgium, Italy, France, and Sweden. [D'Amico et al. \(2019\)](#) conducted research for the ANN energy assessment model on five climate zones. These studies are based on simulated data and use Tsanas and Xifara's dataset for the training of their AI-based prediction models, *i.e.*, machine or deep learning, as well as for testing them.

METHODOLOGY

Sample building

There is currently a rapid construction development of residential buildings in the Qassim region. Accordingly, the Ministry of housing in Saudi Arabia launched a program of 381 villas in Buraydah city and 340 in Unayzah city, all with the same design plan. Since this is a typical new detached house in many towns in the Qassim region, it was selected and used in this study. The house plan is used in the IES<VE> simulation software. The architecture layout of the ground floor and first floor are shown in [Fig. 1](#), while [Table 2](#) provides information of the house envelope construction features.

Modeling in IES<VE>

The IES<VE> simulation software was used to model the house for data generating ([IESVE, 2008](#)). The aim of this phase is to generate the data of the building envelope variables to analyze their effect on the building energy performance. As the building is located in Qassim, Saudi Arabia, the corresponding regional weather data file (epw. format) was imported to the software and used in the simulations. The simulation of design variables was restricted to the house's main spaces subjected to air-conditioning, highlighted in orange in [Fig. 2](#). Other spaces of the house, such as WC, staircase and kitchen, highlighted in blue color in [Fig. 2](#), which are not fully air-conditioned were excluded in the simulation. The specifications of the design variables are provided in [Tables 3](#) and [4](#). All thermal properties for glazing, roof and walls were carefully defined in the IES<VE> simulation software based on their *U*-value ([Table 4](#)), which considered the most effective property that affect the building elements' thermal behavior.

Table 1 A summary of data regarding previous studies in residential building.

References	Building characteristics	Type of energy	Building type	Location	Dataset size
<i>D'Amico et al. (2019)</i>	Wall area Wall <i>U</i> -value Glazing area Glazing <i>U</i> -value	Energy consumption	Residential	U.S. Midwest	973
<i>Cerquitelli, Malmati & Apiletti, 2019</i>			Residential	Athens, Greece	768
<i>Ciulla et al. (2019)</i>			Residential	N/A	
<i>Chen & Tan (2017)</i>	Height		Residential	N/A	
<i>Li et al. (2019)</i>	Relative compactness		Residential	Alicante, Spain	
<i>Le et al. (2019a)</i>	Wall area		Residential	Irvine	
<i>Naji et al. (2016)</i>	Surface area	Heating load	Residential	Greece, Athens	
<i>Ngo (2019)</i>	Roof area		Residential	Athens, Greece	
<i>Kumar, Pal & Singh (2018)</i>	Glazing area distribution	Cooling load	N/A	N/A	
<i>Nilashi et al. (2017)</i>	Glazing area		Residential	Ho Chi Minh City, Viet Nam	
<i>Sadeghi et al. (2020)</i>	Orientation		Residential	NM	
<i>Sharif & Hammad (2019)</i>			Residential	N/A	
<i>Geyer & Singaravel (2018)</i>			Residential	California	
<i>Gao et al. (2019)</i>	Relative compactness	Heating load	N/A	N/A	837
<i>Tien Bui et al. (2019)</i>	Surface Area		Prototype model	Vietnam	
<i>Cecconi, Moretti & Tagliabue, 2019</i>	Wall <i>U</i> -value Wall thickness (5 different walls)	Energy Consumption	Residential	Istanbul	180
<i>Navarro-Gonzalez & Villacampa (2019)</i>	Relative compactness Glazing area Glazing area distribution Roof area Overall Height Orientation Glazing area Glazing area distribution	Heating load Cooling load	Residential Office Others	Athens, Greece	768 ++
<i>Le et al. (2019b)</i>	Insulation K-value Insulation thickness Wall type	Dataset 1:			Two datasets: 180

(Continued)

Table 1 (continued)

References	Building characteristics	Type of energy	Building type	Location	Dataset size
	Relative compactness	Energy Consumption			
	Surface area				+
	Wall area		Residential	Istanbul, Turkey	
	Roof area				768
	Overall height	Dataset 2:			
	Orientation				=
	Glazing area	Cooling load			
	Glazing distribution				948

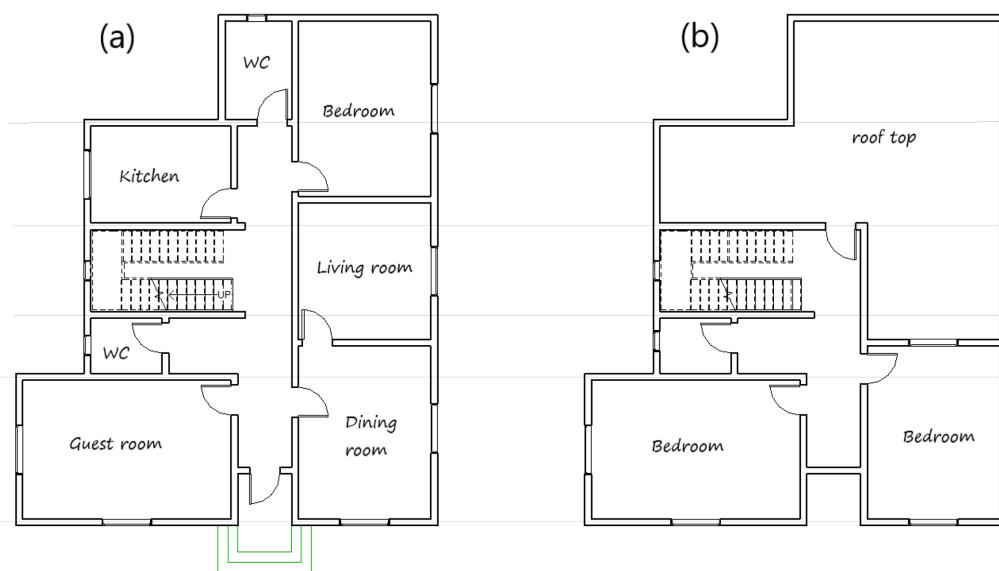


Figure 1 The architecture layout (A) ground floor and (B) first floor of the house sample.

Full-size  DOI: 10.7717/peerj-cs.856/fig-1

Table 2 House envelope construction features.

House features	Description
Location	Buraydah (Coordinates: 26°22'17.8"N 43°51'29.4"E)
Orientation	Front elevation facing South
Shape	Typical Square and Rectangular combination of spaces
Celling Height	3 m
Floor Area	118.1 m ² (Ground Floor); 66.4 m ² (First Floor)
Window Wall Ratio	10–15%
Exterior Walls	15 mm Plaster (Dense) + 10 mm Cement + 200 mm Concrete Block (Medium) + 10 mm Cement + 15 mm Plaster (Lightweight)
Roof	10 mm Ceramic tiles + 30 mm Concrete layer + 10 mm Extruded Polystyrene + 150 mm Reinforced Concrete (Dense)
Windows	4 mm Double clear glass

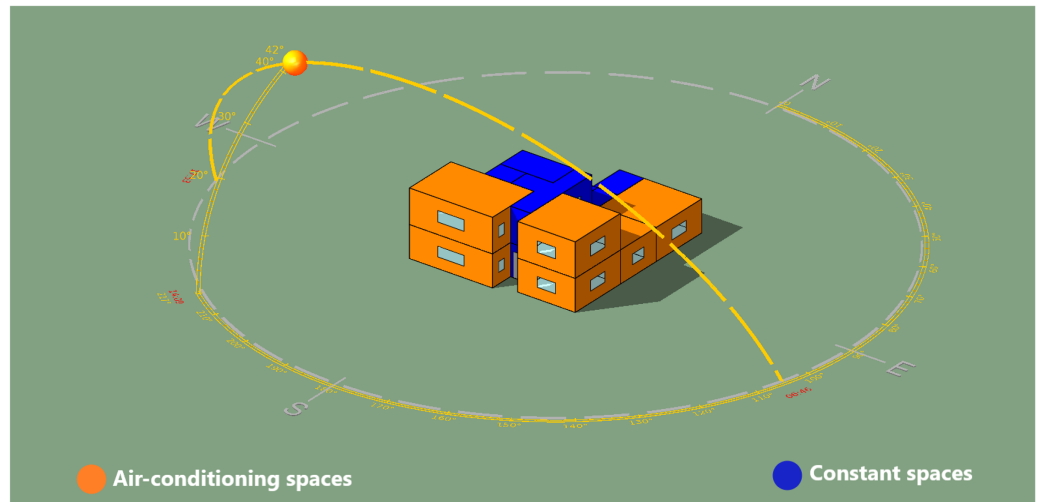


Figure 2 House sample modeling in IESVE with Qassim weather station (Sun-path).

Full-size DOI: 10.7717/peerj-cs.856/fig-2

Table 3 Air-conditioned spaces information generated from IES<VE>.

Space ID	Space name (Real)	Max. height (m)	Volume (m ³)	Floor area (m ²)	Floor perimeter (m)	Ext. wall area (m ²)	Ext. window area (m ²) 10%
SP00000C	Bed room	5.6 (2nd f)	81.026	28.938	21.94	40.945	4.095
SP000000	Bedroom	2.8 (1st f)	65.414	23.362	19.5	27.303	2.73
SP000002	Living Room	2.8 (1st f)	51.614	18.433	17.18	12.183	1.218
SP000003	Dining room	2.8 (1st f)	64.67	23.097	19.38	31.531	3.153
SP000009	Guest room	2.8 (1st f)	81.026	28.938	21.94	40.945	4.095
SP000005	Bed room	5.6 (2nd f)	64.67	23.097	19.38	43.394	4.339
	Total	–	408.42	145.865	–	196.301	19.63

Furthermore, properties of the building elements such as doors, window frame and floors were kept constant in the IES<VE> for the simulation.

Input and output variables

As shown in Table 4, eight different design parameters of a typical house in the Qassim region were considered in order to generate the energy data to predict the whole building energy consumption. Table 4 included descriptions of each design parameters group with possible number of values. All these design parameters and values were applied in the IES<VE> simulation software and the energy consumption values in terms of cooling and heating consumption (output variables), respectively, were obtained as output from the simulation experiment. Building size and floor height have two different values that were constructed in the ModelIT application in the IES<VE> simulation software. The WWR applied to each building size and floor height for the whole external wall that exposed to the outdoor in all directions is also documented in Table 4. The remaining

Table 4 Descriptions of input and output variables in the model simulation.


Features	Description	Variables
Building Size	Spaces in the house subjected to air-conditioning (highlighted in orange color in Fig. 2)	145.86 m ²
		184.53 m ²
Floor Height	This is referred to the internal ceiling height of spaces	2.8 m
		3.0 m
Glazing Area	Net area of windows	23.25 m ²
		24.15 m ²
		69.75 m ²
		72.46 m ²
		116.24 m ²
		120.76 m ²
		162.74 m ²
		169.07 m ²
		209.24 m ²
		217.37 m ²
Wall Area	Net area of walls	217.37 m ²
		209.24 m ²
		169.07 m ²
		162.74 m ²
		120.76 m ²
		116.24 m ²
		72.46 m ²
		69.75 m ²
		24.15 m ²
		23.25 m ²
WWR	Window to wall ratio of all the external wall that exposed to outdoor in all sides	10%
		30%
		50%
		70%
		90%
Win Glazing U-value	Refer to thermal properties of glazing window which calculated by (W/m ² K)	0.97
		1.63
		2.87
		3.23
		4.61
Roof U-value	Refer to thermal properties of the covering of the specified spaces in the model calculated by (W/m ² K)	5.60
		0.13
		0.22
		0.35
		0.47
External Wall U-value	Wall envelope for the specified spaces calculated by (W/m ² K)	0.26

Table 4 (continued)

Features	Description	Variables
		0.34
		0.60
		1.03
		1.62
		2.11
		2.82
		3.34
Cooling Load	Refer to the sensible cooling load through the space's envelope (wall, window and roof) calculated by KWh per year	-
Heating Load	Refer to the sensible heating load through the space's envelope (wall, window and roof) calculated by KWh per year	-

	Building Area (m ²)	Floor Height (m)	Glazing Area (m ²)	Wall Area (m ²)	WWR %	Glazing U-value (W/m ² K)	Roof U-value (W/m ² K)	Wall U-value (W/m ² K)	Cooling (KWh/m ² .yr)	Heating (KWh/m ² .yr)
1										
2	184.53	2.8	23.25	209.24	10	0.97	0.13	3.34	250.25	3.79
3	184.53	2.8	23.25	209.24	10	0.97	0.13	2.82	223.19	1.94
4	184.53	2.8	23.25	209.24	10	0.97	0.13	2.11	184.38	1.20
5	184.53	2.8	23.25	209.24	10	0.97	0.13	1.62	155.23	0.75
6	184.53	2.8	23.25	209.24	10	0.97	0.13	1.03	118.87	0.28
7	184.53	2.8	23.25	209.24	10	0.97	0.13	0.6	89.22	0.05
8	184.53	2.8	23.25	209.24	10	0.97	0.13	0.34	70.63	0.00
9	184.53	2.8	23.25	209.24	10	0.97	0.13	0.26	63.87	0.00
10	184.53	2.8	23.25	209.24	10	0.97	0.22	3.34	255.23	4.10
11	184.53	2.8	23.25	209.24	10	0.97	0.22	2.82	228.34	2.24
12	184.53	2.8	23.25	209.24	10	0.97	0.22	2.11	189.84	1.47
13	184.53	2.8	23.25	209.24	10	0.97	0.22	1.62	160.95	1.00
14	184.53	2.8	23.25	209.24	10	0.97	0.22	1.03	124.97	0.47
15	184.53	2.8	23.25	209.24	10	0.97	0.22	0.6	95.71	0.16
16	184.53	2.8	23.25	209.24	10	0.97	0.22	0.34	77.44	0.04
17	184.53	2.8	23.25	209.24	10	0.97	0.22	0.26	70.83	0.03
18	184.53	2.8	23.25	209.24	10	0.97	0.35	3.34	259.96	4.28
19	184.53	2.8	23.25	209.24	10	0.97	0.35	2.82	233.23	2.41
20	184.53	2.8	23.25	209.24	10	0.97	0.35	2.11	195.03	1.64

Figure 3 A snapshot of our proposed dataset.

Full-size  DOI: 10.7717/peerj-cs.856/fig-3

design parameters based on the U -values were carefully inserted in APACHE application in the IES<VE> simulation software.

As mentioned earlier, all the design parameters were applied to the main spaces only (Table 3) to ensure more reliable and accurate energy data for energy prediction. A total of 3,840 data series were introduced and simulated in the IES<VE> simulation software. A snapshot of our proposed dataset is shown in Fig. 3. Table 5 illustrates the descriptive statistics of the generated data: minimum, maximum, mean, standard deviation, variance, and skewness values.

Table 5 Statistical descriptive of the IES<VE> simulation software generated dataset.

Features	Descriptive index								
	Count	Minimum	Maximum	Mean		Std. deviation	Variance	Skewness	
				Statistic	Std. Error			Statistic	Std. Error
Building Area m ²	3,840	145.86	184.53	165.2	0.31	19.33	373.94	0.0	0.04
Floor Height m	3,840	2.8	3.0	2.9	0.0016	0.10	0.01	0.0	0.04
Glazing Area	3,840	19.63	217.37	110.07	1.018	63.11	3,983.73	0.066	0.04
Wall Area	3,840	19.63	217.37	110.07	1.018	63.11	3,983.73	0.066	0.04
WWR %	3,840	10	90	50.00	0.456	28.288	800.20	0.0	0.04
Win <i>U</i> -value (W/m ² K)	3,840	0.97	5.60	3.16	0.025	1.59	2.56	0.150	0.04
Roof <i>U</i> -value (W/m ² K)	3,840	0.13	0.47	0.29	0.002	0.128	0.017	0.130	0.04
Wall <i>U</i> -value (W/m ² K)	3,840	0.26	3.34	1.51	0.018	1.085	1.179	0.394	0.04
Cooling (KWh/m ² . yr)	3,839	5.45	671.60	336.85	2.18	135.31	18,309.8	0.239	0.04
Heating (KWh/m ² . yr)	3,839	0.0	7.03	0.95	0.02	1.31	1.701	1.892	0.04

Methods and statistical analysis results

This section analyses first the main statistical properties of the variables of the new dataset with the help of histograms and scatterplots. Then, the relationship between the input and output variables is analyzed using the Spearman rank correlation coefficient. Finally, our dataset is analyzed using two machine learning approaches, the Multilayer Regression (MLR) and Multilayer Perceptron (MLP) methods, respectively.

Data exploration

The simulated buildings were generated using the IES<VE> simulation software for Buraydah city. The Qassim province was chosen as it has a hard-arid climate with exceptionally hot summers and cool winters, requiring a lot of energy for cooling and heating residential buildings. The dataset is available at [Almhafdy \(2021\)](#) and contains 3,840 records. The following nine constant characteristics were used: location (Buraydah), orientation (front façade oriented to south), shape (rectangular and square spaces), ceiling height (3 m), floor area (ground floor 118.1 m²; first floor 66.4 m²), window wall ratio (10–15%), Exterior walls (0.015 m plaster + 0.01 mm cement + 0.020 m concrete block (medium) + 0.01 m cement + 0.015 m plaster (lightweight), roof (0.01 m ceramic tiles + 0.03 m concrete layer + 0.01 m extruded polystyrene + 0.015 m reinforced concrete, and windows (0.004 m double clear glass).

Two building sizes were used 145.86 m² and 184.53 m². For each building size two floor heights of 2.8 m and 3 m were used; five different WWR as percentage of all external wall exposed to outdoor were used: 10%, 30%, 50%, 70%, and 90%; six win-value were simulated: 0.97, 1.63, 2.87, 3.23, 4.61, and 5.60); four different roof *U*-value were simulated: 0.13, 0.22, 0.35, and 0.47; and eight wall *U*-value were applied to each roof *U*-value. This is illustrated in [Fig. 4](#).

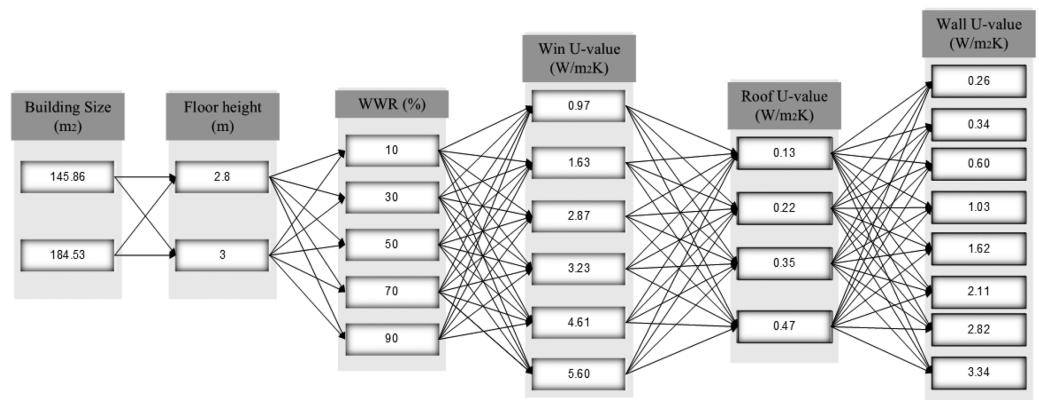


Figure 4 Input design parameters groups for energy consumption of building.

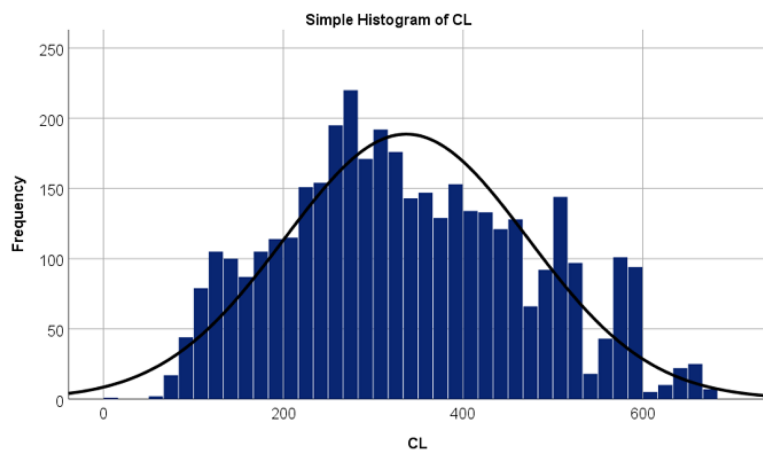
Full-size DOI: 10.7717/peerj-cs.856/fig-4

Table 6 Mathematical representation of the input and output variables with the number of possible values.

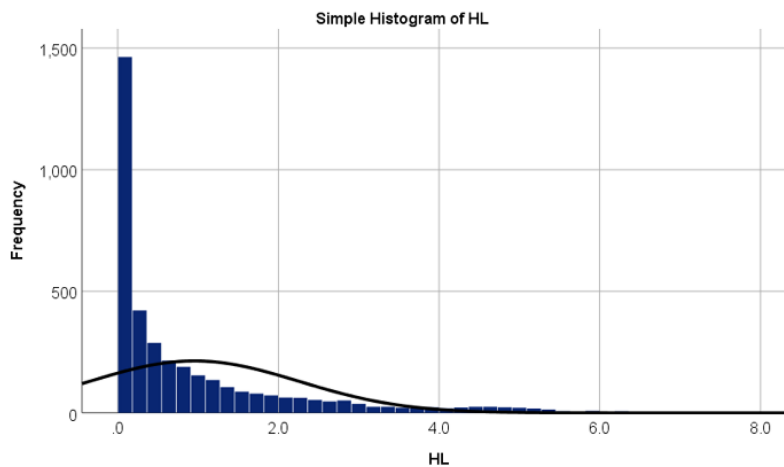
Mathematical representation	Input or output variable/Feature	No. of possible values	Label for charts
I1	Building Size	2	BA
I2	Floor Height	2	FH
I3	Glazing Area	10	GA
I4	Wall Area	10	WA
I5	WWR	5	WWR
I6	Win Glazing U -value	6	WinU
I7	Roof U -value	4	RU
I8	External Wall U -value	8	WU
O1	Cooling Load	3,659	CL
O2	Heating Load	2,674	HL

Accordingly, we obtained $2 * 2 * 5 * 6 * 4 * 8 = 3,840$ building samples. The simulated buildings are characterized by eight building features (input variables), and their output HL and CL were recorded, as summarized in Table 6.

Statistical properties of the variables were first analyzed with visualization of the empirical probability distributions of all the input and output variables (Tsanas & Xifara, 2012). These are provided in Fig. 5 which presents the probability density estimates using histograms of the output variable: the cooling load and the heating load. Figure 5A shows the frequency distribution for the cooling load output variable that resulted in the 3,840 records in the dataset and it describes that the most values are within a range of 100 to 600. While in Fig. 5B, the frequency distributions show that most of the values of the output variable heating load are ranged between 0.0 to 0.2. As a result, the necessity to experiment with machine learning approaches such as multiple linear regression (MLR) and multilayer perceptron (MLP) is intuitively justified.



(a)



(b)

Figure 5 Probability density estimates using histograms of the output variable (A) cooling load, and (B) heating load.

Full-size  DOI: [10.7717/peerj-cs.856/fig-5](https://doi.org/10.7717/peerj-cs.856/fig-5)

Statistical analysis

Due to the general non-Gaussian nature of the data, the Spearman rank correlation coefficient was used to derive a statistical metric for the strong relationship between each input variable with each of the two output variables (*Tsanas & Xifara, 2012*), which is given in [Table 7](#). It is evident that several of the input variables are highly associated, such as GA (Glazing Area) and WWR (Window to Wall Ratio). As it is naturally expected, the variables GA and WWR are almost inversely proportional to WA.

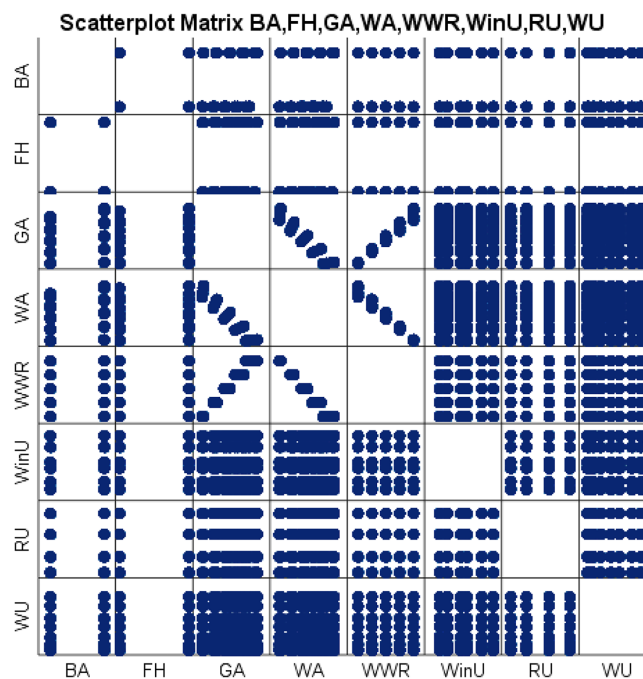
We can similarly depict the bivariate correlations between the eight input variables using a scatter plot matrix. A scatter plot matrix is a grid (or matrix) that represents a single view with multiple scatterplots in a matrix format (*Elmqvist, Dragicevic & Fekete, 2008*). Each scatter plot in the matrix depicts the relationship between two variables, allowing for the exploration of multiple relationships in a single graph. [Figure 6](#) shows a scatter plot matrix of our eight input variables. The position of each dot on the horizontal

Table 7 Correlations matrix using Spearman rank correlation between the eight input variables.

	BA	FH	GA	WA	WWR	WinU	RU	WU
BA	1.000	0.000	0.173	0.173	0.000	0.000	0.000	0.000
FH	0.000	1.000	0.087	0.087	0.000	0.000	0.000	0.000
GA	0.173	0.087	1.000	-0.925	0.981	0.000	0.000	0.000
WA	0.173	0.087	-0.925	1.000	-0.981	0.000	0.000	0.000
WWR	0.000	0.000	0.981	-0.981	1.000	0.000	0.000	0.000
WinU	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
RU	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
WU	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Note:

BA: Building Size, FH: Floor Height, GA: Glazing Area, WA: Wall Area, WWR: Window to Wall Ratio, WinU: Win Glazing *U*-value, RU: Roof *U*-value, WU: External Wall *U*-value.

**Figure 6** Scatter plot matrix representation of the eight input variables.

Full-size DOI: 10.7717/peerj-cs.856/fig-6

and vertical axis indicates values for an individual data point. For each pairwise combination of variables chosen, a scatter plot is constructed.

Machine learning-based analysis

The main objective of this study is to describe a dataset generated for the energy consumption of buildings in the arid climate. This section makes use of two machine learning models, namely Multiple Linear Regression (MLR) and Multilayer Perceptron (MLP). These two models were chosen to examine the viability of the developed dataset in predicting the buildings energy consumption in terms of cooling and heating loads. In a

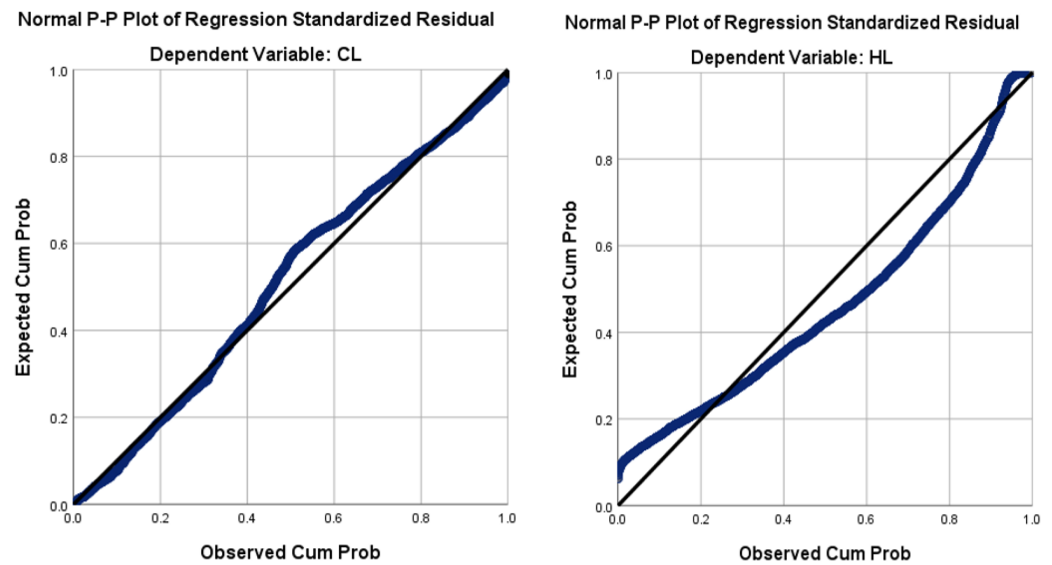


Figure 7 The normal P-P plot of the regression standardized residual for our dependent variables CL and HL. [Full-size !\[\]\(5f471a71b78d7676bc356df190b88ab4_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-7](https://doi.org/10.7717/peerj-cs.856/fig-7)

recent study of ours (*Al-Shargabi et al., 2021*), we applied deep learning and created various models to predict the energy consumption of buildings using the dataset described in this study.

Multiple linear regression analysis

Multiple regression extends simple linear regression to predict the value of a variable (the outcome, target or criterion variable) based on the values of two or more other variables (the predictor, explanatory or regressor variables) (*Tian et al., 2017*).

This section examines the distribution of the output variables (CL and HL) using the normal P-P plot, and the scatter plot of the regression standardized residual. The normal P-P plot of the standardized residual for dependent variables CL and HL is shown in *Fig. 7*, which corroborates that CL is normally distributed while HL is not.

Cross validation (CV) is a common statistical re-sampling technique used in this paper. The dataset is divided into two subsets: a training subset and a testing subset. The training subset is used to derive model parameters, while the testing subset is used to compute errors (out-of-sample error or testing error). In particular, 10-fold CV (*Uyank & Güler, 2013*) is used as the learner testing method. We investigate how accurate the actual statistical mapping is reporting out-of-sample errors after conducting the exploratory statistical analysis, which provides important insight into the strength of the association between the input parameters and the output variables. The mean value of each MLR coefficient over the 10-fold CV iterations is obtained and used for predicting CL and HL in *Eqs. (1)* and *(2)*, respectively.

$$MLR_{CL} = 11.448 - 3.24I_1 - 75.083I_2 + 2.468I_3 + 3.313I_4 + 5.519I_5 + 34.84I_6 + 37.093I_7 + 29.89I_8 \quad (1)$$

$$MLR_{HL} = -0.029 - 0.035I_1 - 1.209I_2 + 0.035I_3 + 0.035I_4 + 0.004I_5 + 0.401I_6 + 0.795I_7 + 0.508I_8 \quad (2)$$

Multilayer perceptron analysis

In this model, using our proposed dataset, an ANN using the Multilayer perceptron method, which is one of the most commonly used methods for building an ANN ([Hastier, Tibshirani & Friedman, 2009](#)), is built in SPSS.

Artificial neural networks (ANN) are nonlinear models that fall into the artificial intelligence technique category known as black-box models ([Heddam, 2016](#)). The multilayer perceptron neural network (MLP) ([Rumelhart, Hinton & Williams, 1985](#)) is one of the most extensively used ANN architectures in the literature, and it is extensively employed in hydrological, water resources, and environmental applications. Three layers make up the MLP: the input layer contains the independent variables, the output layer contains the dependent variable, and one or more hidden layers may also be present. The parameters of the MLP model are its weights and biases. It was used to alter the weights and biases of the training subsets, and the MLP was then trained with random beginning values. To choose the model with the lowest MSE between actual and predicted CL and HL, the training process is repeated many times. Neural networks with Sigmoid activation functions in their hidden layers and linear activation functions in their output layers, commonly known as the identity function, are employed for this research.

To select the number of hidden layers, automatically architecture selection is chosen. The following three different distributions for the dataset are applied: (i) 70% to train the NN and 30% to test the NN; (ii) 80% to train the NN and 20% to test the NN; (iii) 90% to train the NN and 10% to test the NN. [Figures 8 and 9](#) show the obtained NNs to predict CL and HL from the set of 8 input variables, respectively.

The importance score of each of the eight independent variables in the prediction of each of the output variables is computed and given as [Table 8](#). According to [Table 8](#), the top five important input variables when predicting both the CL and HL output variables are WWR, WinU, GA, WA, and WU. [Figures 10 and 11](#) shows the importance distribution percentages of the input variables as determined by the MLP for the CL and HL output variables, respectively. The top five important input variables are further investigated in terms of their effect on predicting buildings energy consumption in “Concluding Remarks and Future Research Directions”. These five input variables are the base to create various combinations to several prediction models of the CL and HL.

Error and performance measures

This section reports on the general performance of the trained methods that were discussed in the previous section. The models are compared using three performance measures, namely, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (R^2).

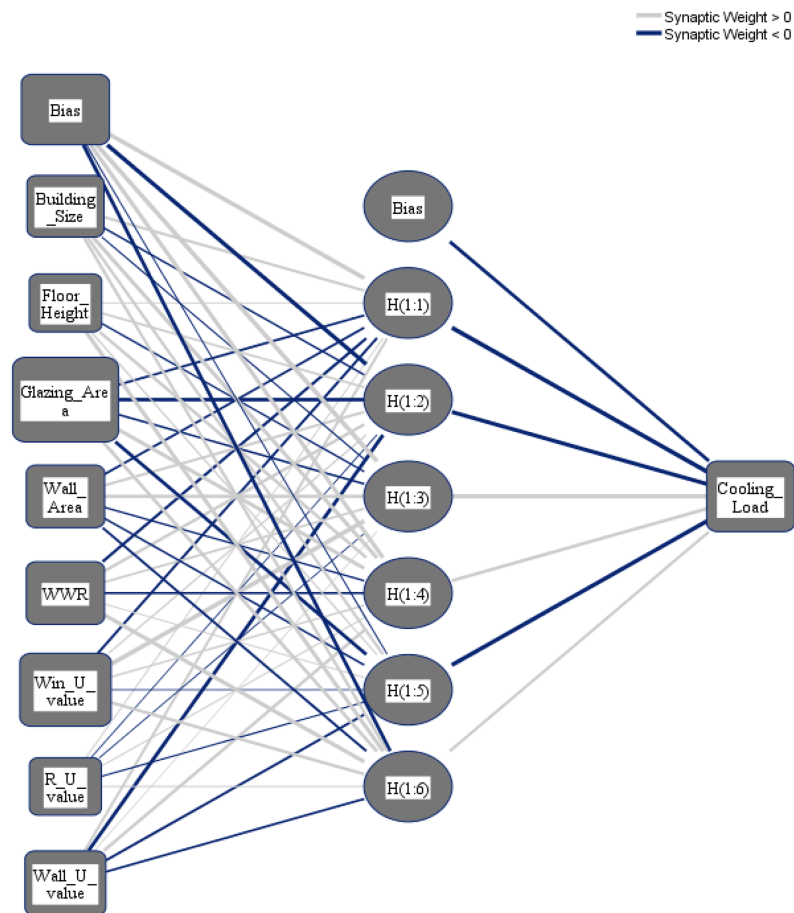


Figure 8 Multilayer perceptron model for predicting the cooling load output from the input variables. [Full-size !\[\]\(1679558f37f6db0dd8360a2a7e913e90_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-8](https://doi.org/10.7717/peerj-cs.856/fig-8)

The average difference between expected and actual variables, such as heating and cooling loads, is known as the Mean Absolute Error (MAE). In (Eq. 3), the following equation demonstrates how MAE is calculated:

$$MAE = (1/n) \times \sum_{i=1}^n |p_i - y_i| \quad (3)$$

Prediction errors are calculated by calculating the Root Mean Square Error (RMSE). Large variations between expected and actual results can be captured using this method. The lower the RMSE, the more accurate the model is. In (Eq. 4), the RMSE is determined using the following equation:

$$RMSE = \sqrt{(1/n) \times \sum_{i=1}^n [p_i - y_i]^2} \quad (4)$$

The coefficient of determination (R^2) indicates how much of the variance in the dependent variable can be predicted using the independent variables, such as heating and

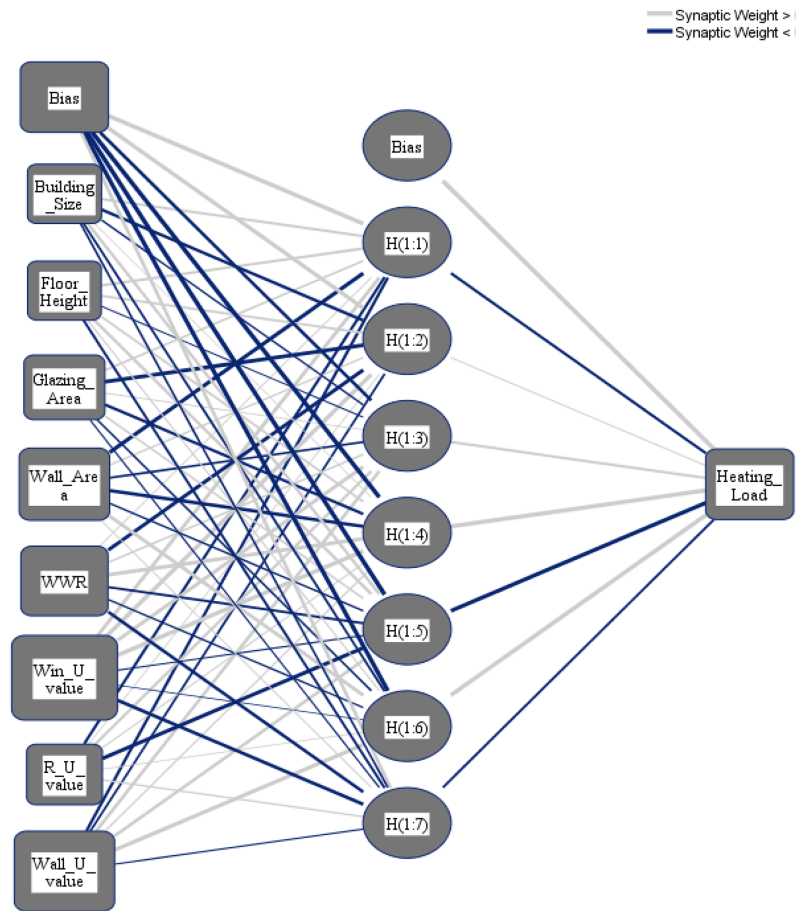


Figure 9 Multilayer perceptron model for predicting the heating load output from the input variables. [Full-size !\[\]\(fcc3264021d438d9732560e78099f674_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-9](https://doi.org/10.7717/peerj-cs.856/fig-9)

Table 8 Importance of the input variables as determined by the MLP for the output variables.

Measure	Importance score with CL	Importance score with HL
BA	0.049 ± 0.015	0.067 ± 0.025
FH	0.024 ± 0.003	0.023 ± 0.009
GA	0.209 ± 0.129	0.087 ± 0.033
WA	0.126 ± 0.028	0.111 ± 0.040
WWR	0.240 ± 0.142	0.157 ± 0.031
WinU	0.230 ± 0.041	0.296 ± 0.009
RU	0.015 ± 0.002	0.038 ± 0.002
WU	0.108 ± 0.018	0.252 ± 0.009

cooling loads. The closer value to 1, the higher performance model and the stronger relationship, as calculated in (Eq. 5).

$$R^2 = \frac{\sum_{i=1}^n (p_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

Importance with Cooling Load

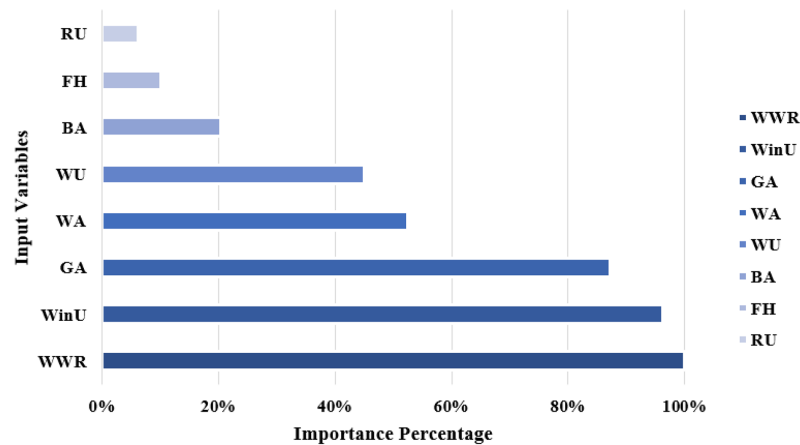


Figure 10 Importance distribution of the input variables as determined by the MLP for the cooling load output variables. [Full-size !\[\]\(95c552df6353b48e62ab71c0e20270ca_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-10](https://doi.org/10.7717/peerj-cs.856/fig-10)

Importance with Heating Load

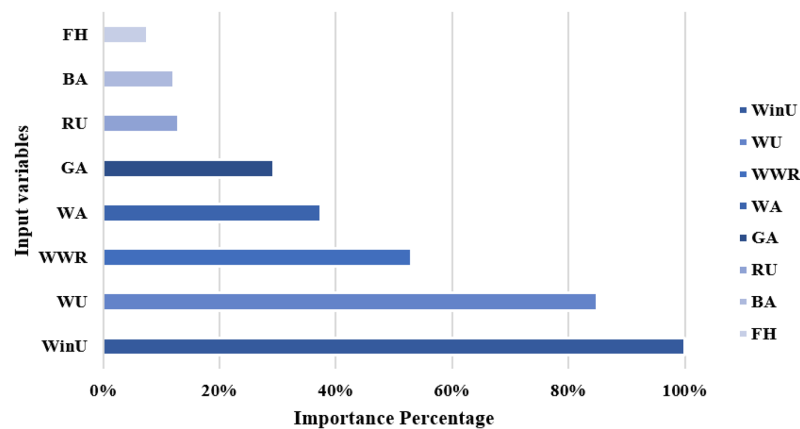


Figure 11 Importance distribution of the input variables as determined by the MLP for the heating load output variables. [Full-size !\[\]\(0f04eac0f44b45c27c855fb13a3dc0b4_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-11](https://doi.org/10.7717/peerj-cs.856/fig-11)

where p_i identifies the predicted value for sample i , y_i identifies the actual value for sample i , n is the sample size, \bar{y} indicates the mean of the predicted values.

RESULTS AND DISCUSSIONS

This study investigated various combinations of the eight building characteristics variables as inputs to the MLP and MLR models in order to examine the effect of these variables on the energy consumption in terms of heating and cooling loads. During this research, a total of eight different models were created and compared (Tables 9 and 10).

According to testing data, the MAE, RMSE, and R^2 statistics of several MLP and MLR models in predicting the cooling load (CL) are shown in Table 9. Table 9 shows significant differences across the eight MLP models based on the three performance indicators.

Between 21.78 and 23.2 (MAE, RMSE, and R^2), respectively, the values of MAE, RMSE,

Table 9 Out of sample MAE, RMSE, and R^2 for predicting the CL output variable for the MLR and MLP models.

Model		Input variables	MAE	RMSE	R^2
MLP	M1	WinU+WWR+WU+GA+WA	23.2	42.92	0.999
	M2	WinU+WWR+WU+GA+WA+BA	23.05	41.54	0.999
	M3	WinU+WWR+WU+GA+WA+FH	22.71	38.69	0.997
	M4	WinU+WWR+WU+GA+WA+RU	22.88	40.14	0.998
	M5	WinU+WWR+WU+GA+WA+BA+FH	22.39	35.7	0.995
	M6	WinU+WWR+WU+GA+WA+BA+RU	22.51	37.22	0.996
	M7	WinU+WWR+WU+GA+WA+FH+RU	22.07	32.56	0.993
	M8	All: WinU+WWR+WU+GA+WA+BA+FH+RU	21.78	29.123	0.992
MLR	M1	WinU+WWR+WU+GA+WA	47.91	66.32	0.990
	M2	WinU+WWR+WU+GA+WA+BA	46.97	61.40	0.984
	M3	WinU+WWR+WU+GA+WA+FH	47.37	63.56	0.986
	M4	WinU+WWR+WU+GA+WA+RU	47.62	64.87	0.988
	M5	WinU+WWR+WU+GA+WA+BA+FH	46.26	57.43	0.979
	M6	WinU+WWR+WU+GA+WA+BA+RU	46.66	59.70	0.982
	M7	WinU+WWR+WU+GA+WA+FH+RU	46.38	58.15	0.980
	M8	All: WinU+WWR+WU+GA+WA+BA+FH+RU	46.020	56.015	0.978

Table 10 Out of sample MAE, RMSE, and R^2 for predicting the HL output variable for the MLR and MLP models.

Model		Input variables	MAE	RMSE	R^2
MLP	M1	WinU+WWR+WU+GA+WA	0.180	0.376	1
	M2	WinU+WWR+WU+GA+WA+BA	0.177	0.346	1
	M3	WinU+WWR+WU+GA+WA+FH	0.179	0.368	1
	M4	WinU+WWR+WU+GA+WA+RU	0.175	0.333	0.996
	M5	WinU+WWR+WU+GA+WA+BA+FH	0.174	0.320	0.992
	M6	WinU+WWR+WU+GA+WA+BA+RU	0.170	0.284	0.981
	M7	WinU+WWR+WU+GA+WA+FH+RU	0.172	0.308	0.989
	M8	All: WinU+WWR+WU+GA+WA+BA+FH+RU	0.167	0.260	0.433
MLR	M1	WinU+WWR+WU+GA+WA	0.955	1.567	0.469
	M2	WinU+WWR+WU+GA+WA+BA	0.942	1.455	0.521
	M3	WinU+WWR+WU+GA+WA+FH	0.948	1.510	0.493
	M4	WinU+WWR+WU+GA+WA+RU	0.945	1.481	0.507
	M5	WinU+WWR+WU+GA+WA+BA+FH	0.935	1.399	0.552
	M6	WinU+WWR+WU+GA+WA+BA+RU	0.921	1.269	0.627
	M7	WinU+WWR+WU+GA+WA+FH+RU	0.928	1.337	0.587
	M8	All: WinU+WWR+WU+GA+WA+BA+FH+RU	0.915	1.223	0.656

and R^2 were found. MAE and RMSE performance metrics have the lowest values when all eight of the building's identifying attributes are supplied into the M8 model (WinU, WWR, WU, GA, WA, BA, FH, and RU). The highest R^2 values were found with the M1 and

M2 models, however, the M8 model still had the highest value. As can be seen from the results, the MLP M8 model has excellent cooling load (CL) performance and outstanding overall accuracy in predicting cooling load.

Table 9 also displays the results of the cooling load (CL) prediction using MLR models based on the testing data. The MAE and RMSE metrics based on MLR models yield poorer results than those based on MLP models. Furthermore, the eight MLR models revealed considerable variances depending on the three performance measurements criterion, as shown in Table 9. MAE, RMSE, and R^2 values varied from 46.02 to 47.91, 56.01 to 66.32, and 0.978 to 0.99, respectively. The M8 model, which employs all eight building characteristics variables as input, also yields the lowest values of the MAE and RMSE performance measures (WinU, WWR, WU, GA, WA, BA, FH, and RU). The highest values for the R^2 measure were obtained with the M1 model, which was not far off from the value obtained with the M8 model. In terms of MAE, RMSE, and R^2 statistics, Table 9 compares the effectiveness of several MLP and MLR models in forecasting cooling load (CL).

Similarly, Table 10 reported the results obtained in predicting the heating load (HL) based on the same three performance measures. The MAE, RMSE, and R^2 values for the MLP models ranged from (0.167 to 0.18), (0.26 to 0.37), and (0.43 to 1.00), respectively, according to Table 10. The M8 model, which employs all eight building characteristics variables as input, likewise produces the lowest MAE and RMSE performance scores (WinU, WWR, WU, GA, WA, BA, FH, and RU). With the M1, M2, and M3 models, the highest R^2 values were found. The MAE and RMSE figures indicate that the MLP model's performance is extremely good, and the MLP M8 model generally achieves good forecast accuracy of heating load (HL). Table 10 also displays the heating load (HL) prediction results derived using MLR models based on the testing data. The MAE and RMSE values based on the MLR models are lower than those based on the MLP models, as evidenced by the cooling load projections in Table 9. Table 10 shows that the eight MLR models differed significantly based on the three performance measurements criterion. MAE, RMSE, and R^2 values varied between (0.915 to 0.955), (1.223 to 1.567), and (0.469 to 0.656), respectively. The M8 model, which employs all eight building characteristics variables as input, also yields the lowest MAE and RMSE values and the highest value of R^2 performance metrics (WinU, WWR, WU, GA, WA, BA, FH, and RU).

In comparison, the prediction accuracy of heating load (HL) for the regression models was higher than the prediction accuracy of cooling load (HL) in both MLP and MLR models for all eight generated combinations, according to the data provided in Table 9.

The comparison of the models was based on graphical plots as scatter plots, box plots, violin plots, and Taylor diagram plots. Figures 12 and 13 show the scatterplots of the actual and the predicted values of the cooling load and the heating loads output variables obtained by MLP and MLR when using all the inputs, as represented in model M8 in Table 9 and 10. The best cooling load results of R^2 with 0.976 was achieved by MLP, whereas the MLR model provides R^2 with 0.839. similarly, the R^2 value for the heating load using MLP model is 0.958 which is better than the 0.438 R^2 value given by the MLR model.

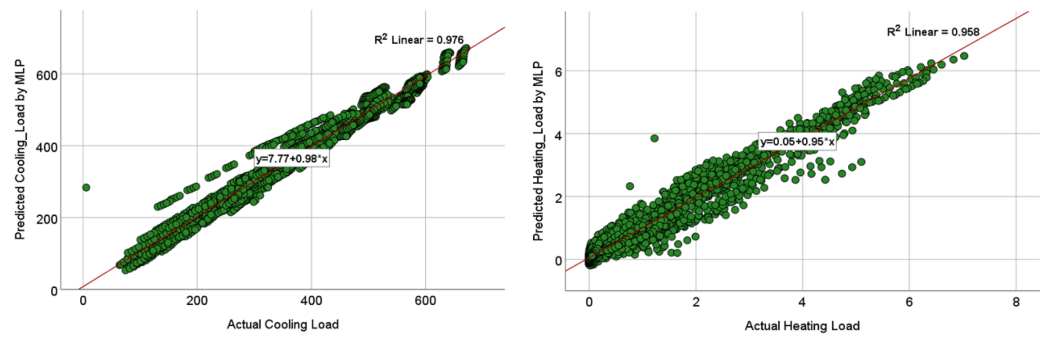


Figure 12 Scatterplots showing the relation between the actual and the predicted values of the cooling load (CL) and heating loads (HL) variables for the MLP M8 model.

Full-size DOI: [10.7717/peerj-cs.856/fig-12](https://doi.org/10.7717/peerj-cs.856/fig-12)

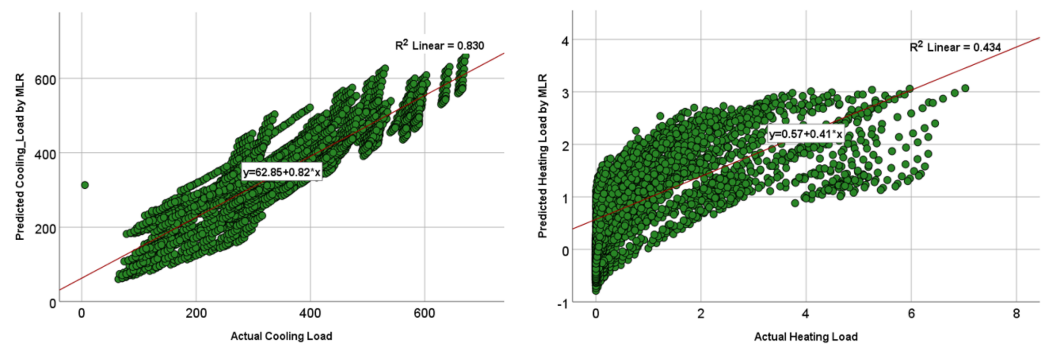


Figure 13 Scatterplots showing the relation between the actual and the predicted values of the cooling load (CL) and heating loads (HL) variables for the MLR M8 model.

Full-size DOI: [10.7717/peerj-cs.856/fig-13](https://doi.org/10.7717/peerj-cs.856/fig-13)

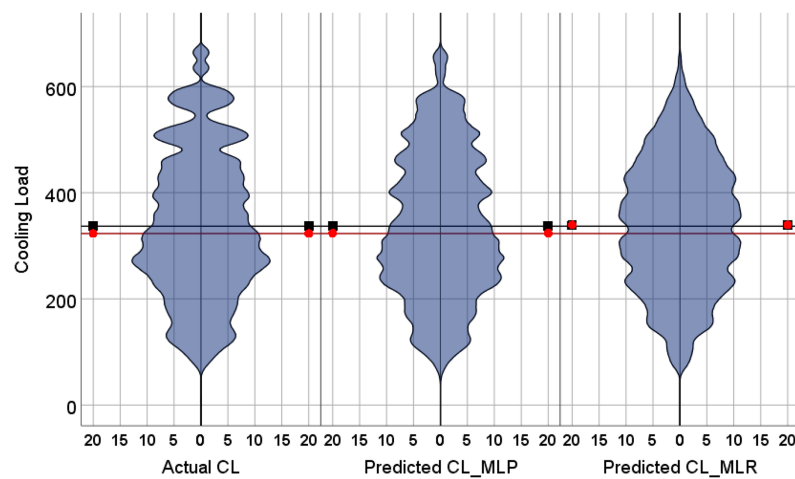


Figure 14 Violin plots of the actual and the predicted values of the cooling load (CL) values obtained by MLP and MLR.

Full-size DOI: [10.7717/peerj-cs.856/fig-14](https://doi.org/10.7717/peerj-cs.856/fig-14)

The violin plots and the box plots for the actual and the predicted values of the heating load and cooling load output variables are illustrated in Figs. 14–17. As in the violin plots presented in Fig. 14, the two lines with a black square and red circle color display the

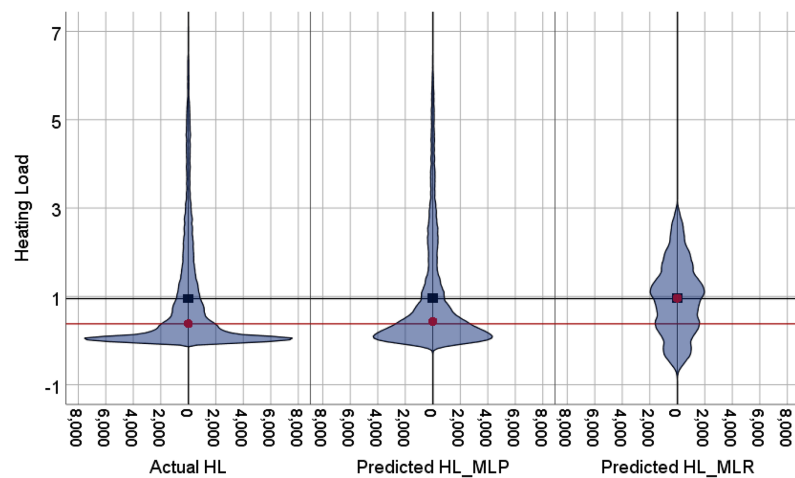


Figure 15 Violin plots of the actual and the predicted values of the heating load (HL) values obtained by MLP and MLR. [Full-size](#) DOI: 10.7717/peerj-cs.856/fig-15

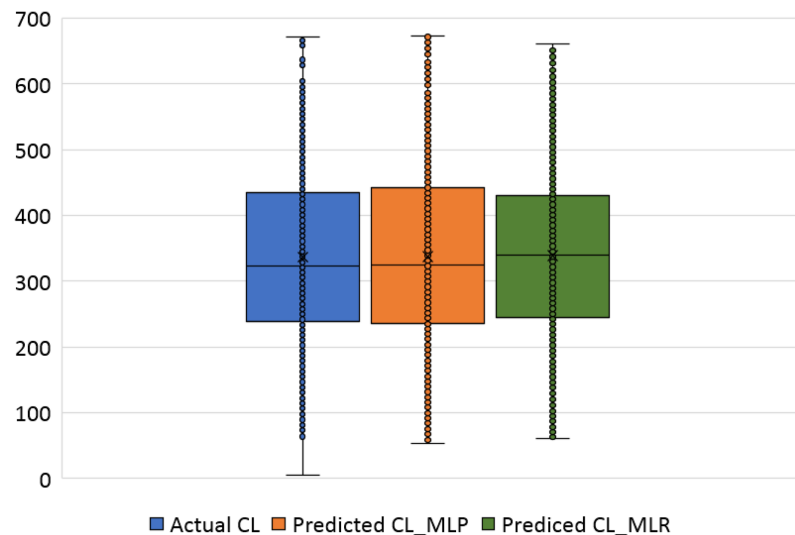


Figure 16 Box plots of the actual and the predicted values of cooling load (CL) values obtained by MLP and MLR. [Full-size](#) DOI: 10.7717/peerj-cs.856/fig-16

mean and the median values of the heating and cooling loads, respectively. The high resemblance between the actual and the predicted heating load was achieved by MLP, especially on the median (323.18 and 323.85), while the MLR median value is 339.36. While the values of the mean for the CL in the actual, predicted MLP and MLR are very close (336.85, 337.08, and 338.75), as illustrated in Table 11. Similarly, for the heating load in Fig. 15 and Table 11, the high similarity between the actual and the predicted heating load was also accomplished by MLP with median values 0.38 and 0.43 where the median of the MLR is 0.96.

Figure 16 illustrates the box plots of the actual and the predicted cooling load by MLP and MLR models. The median is represented by the central line with values 323.17, 323.84, and 339.35 for the actual, the predicted MLP, and the predicted MLR, respectively.

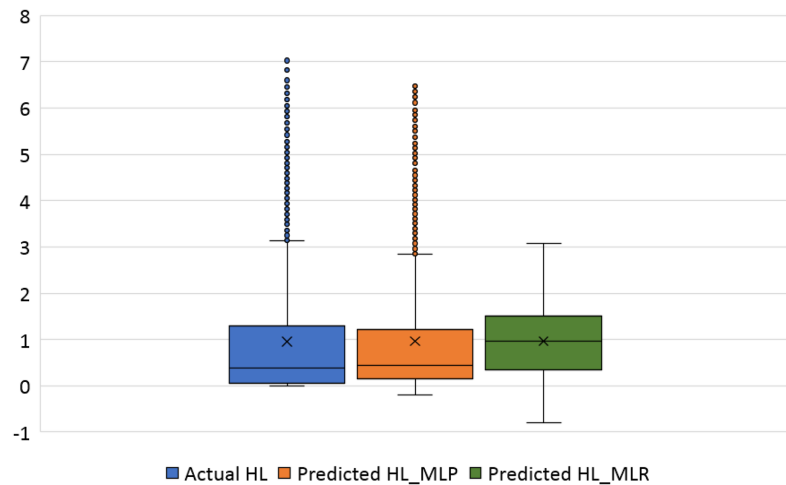


Figure 17 Box plots of the actual and the predicted values of heating load (HL) values obtained by MLP and MLR. [Full-size !\[\]\(fcc3264021d438d9732560e78099f674_img.jpg\) DOI: 10.7717/peerj-cs.856/fig-17](https://doi.org/10.7717/peerj-cs.856/fig-17)

Table 11 The mean and median values obtained by the actual, the MLP, and the MLR predicted models for the CL and HL variables derived from the violin and box plots.

Output variable	Model	Mean		Median	
		Violin plot	Box plot	Violin plot	Box plot
CL	Calculated (simulated)	336.85	336.84	323.18	323.17
	MLP	337.08	337.08	323.85	323.84
	MLR	338.75	338.75	339.36	339.35
HL	Calculated (simulated)	0.95	0.953	0.38	0.383
	MLP	0.96	0.961	0.43	0.433
	MLR	0.96	0.962	0.96	0.96

This indicates that the MLP model is better than the MLR model, as shown in [Table 11](#). The 25th and 75th percentiles are represented by the box's two edges, and the x symbol represents the mean points which have values 336.85, 337.08, and 338.75 for the actual, the predicted MLP, and the predicted MLR, respectively. Likewise, [Fig. 17](#) demonstrates the box plots of the actual and the predicted heating load variables obtained by MLP and MLR models. The median is represented by the central line with values 0.383, 0.433, and 0.96 for the actual, the predicted MLP, and the predicted MLR, respectively. It is clear from the box plots that the MLR model gives better values near the actual cooling and heating loads.

Finally, the Taylor diagram plot was used to compare the MLP and the MLR models for the cooling load and the heating load as in [Figs. 18](#) and [19](#), respectively. Taylor diagram plot is one of the most and highly recommended diagrams for performance comparisons of machine learning ([Zhu et al., 2019](#)). It exhibits three specific statistics: Pearson correlation (R), ratio value, and the normalized standard deviation. The ratio value means the ratio of the normalized variances indicates the relative amplitude of the model and observed variations. It is shown from the two figures that MLP performed

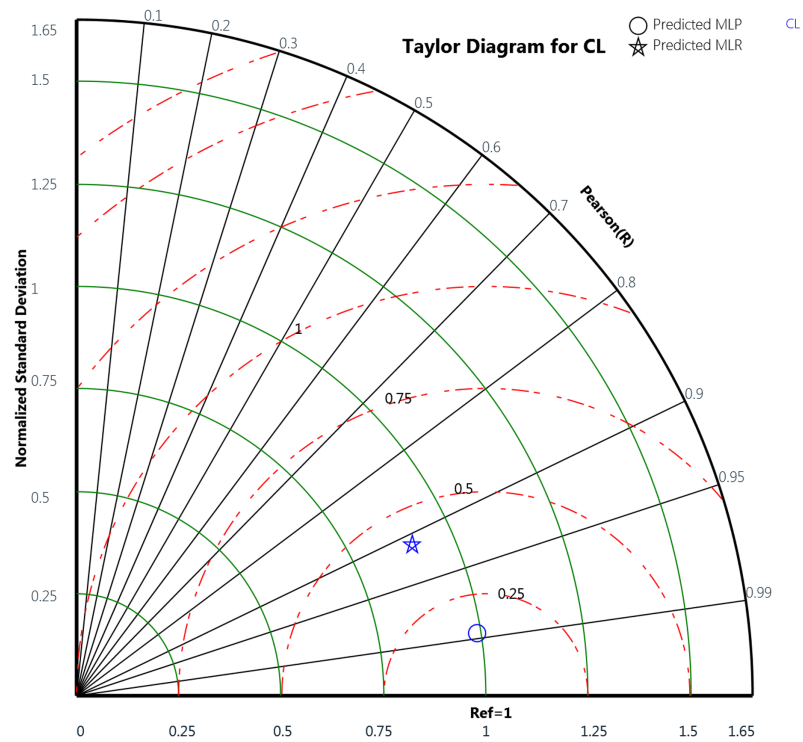


Figure 18 Taylor diagram of the actual and the predicted cooling load (CL) values obtained by MLP and MLR.

Full-size DOI: [10.7717/peerj-cs.856/fig-18](https://doi.org/10.7717/peerj-cs.856/fig-18)

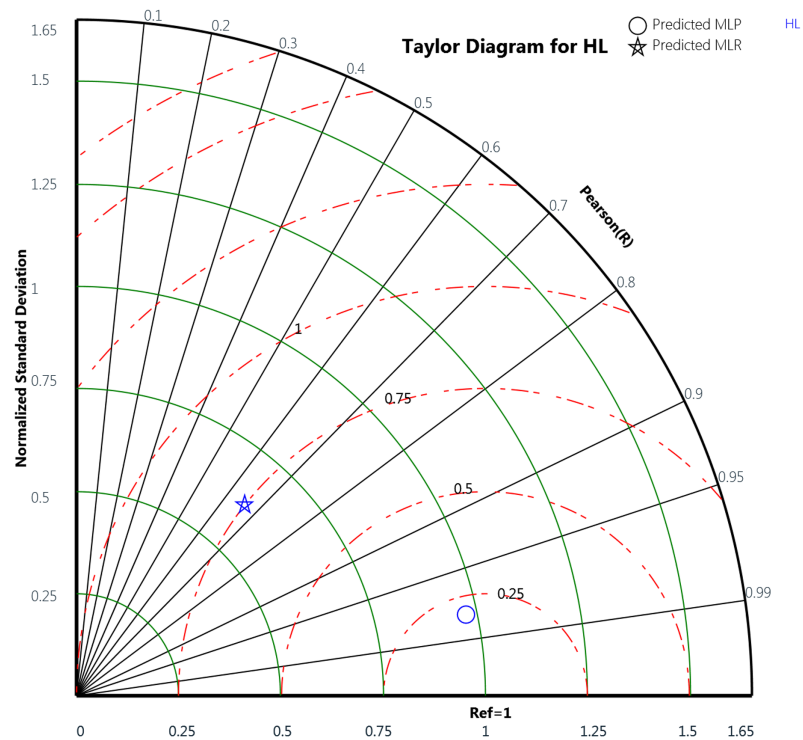


Figure 19 Taylor diagram of the actual and the predicted heating load (HL) values obtained by MLP and MLR.

Full-size DOI: [10.7717/peerj-cs.856/fig-19](https://doi.org/10.7717/peerj-cs.856/fig-19)

Table 12 The ratio and correlation values obtained by the MLP and MLR models for the CL and HL variables using the Taylor diagram.

Output variable	Model	Ratio value	Correlation value
CL	MLP	0.989	0.988
	MLR	0.899	0.911
HL	MLP	0.971	0.979
	MLR	0.622	0.659

better than MLR. In general, the MLP points, represented by the blue circle, are closer to the reference points than the blue star symbols that signify the MLR. The ratio values for the CL variable predicted by the MLP model is 0.989 and the MLR model is 0.899. Whereas in the HL variables, the predicted MLP and the predicted MLR model gives ratio values with 0.971 and 0.622, respectively, as illustrated in Table 12. The table also represents the correlation values of the two MLP and MLR models for the CL and HL variables where the CL the MLP gives 0.988 correlation value while the MLR gives 0.911. For the HL, the MLP and the MLR correlations values are 0.979 and 0.659, respectively. These plots demonstrate that the MLP model predicts the cooling load and the heating load output variables in a better way compared to the MLR model when comparing the actual values with the predicted values.

CONCLUDING REMARKS AND FUTURE RESEARCH DIRECTIONS

Predicting building energy consumption is critical for achieving energy efficiency and sustainability. Nowadays, building energy simulation software is frequently used to assess or predict building energy usage to aid in the design and operation of energy-efficient buildings. This paper investigated the impact of eight input variables on residential buildings heating load (HL) and cooling load (CL), respectively. A variety of classical and non-parametric statistical analytic tools were used to find the most strongly associated input variables with each of the output variables. Then, using the performance measures Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2), two machine learning statistical methods to estimate HL and CL were compared: Multiple linear regression (MLR) and Multilayer perceptron (MLP). Simulation experiments on 3,840 different residential buildings showed that HL and CL can accurately be predicted using the IES<VE> simulation software actual data with low MAE, RMSA, and R^2 values, especially when using the MLP approach.

The findings of this study suggest that predicting building parameters using machine learning methods is a practical and accurate method. Among the major findings of this study is that the MLP models are more accurate in predicting both cooling and heating loads of the buildings, as compared to the MLR models. Also, the best performed MLP model was the one that uses the eight input variables.

Based on the eight buildings characteristics input variables, many various combinations can be created for predicting the energy consumption, however, and due to the time

limitation, only eight combinations have been considered with a focus on the most important input variables.

The obtained results in this paper suggest that future research on the application of additional machine learning and deep learning models to analyze our proposed dataset and comparison with other benchmark datasets is worth considering.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Qassim University, represented by the Deanship of Scientific Research, provided the financial support for this research under the number (coc-2019-2-2-I-5422) during the academic year 1440 AH/2019 AD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Qassim University: coc-2019-2-2-I-5422.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Dina M. Ibrahim conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Abdulbasit Almhafdy conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Amal A. Al-Shargabi conceived and designed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Manal Alghieth analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Ahmed Elragi performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Francisco Chiclana analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at GitHub: <https://github.com/Dr-Dina-M-Ibrahim/A-dataset-for-residential-buildings-energy-consumption-with-statistical-and-machine-learning-analysi.git>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.856#supplemental-information>.

REFERENCES

- Al-Shargabi AA, Almhafdy A, Ibrahim DM, Alghieth M, Chiclana F. 2021. Tuning deep neural networks for predicting energy consumption in arid climate based on buildings characteristics. *Sustainability* **13**(22):12442 DOI 10.3390/su132212442.
- Almhafdy A. 2021. A dataset for residential buildings energy consumption with statistical and machine learning analysis. GitHub. Available at <https://github.com/Dr-Dina-M-Ibrahim/A-dataset-for-residential-buildings-energy-consumption-with-statistical-and-machine-learning-analysi> (accessed 9 December 2021).
- Cecconi FR, Moretti N, Tagliabue LC. 2019. Application of artificial neural network and geographic information system to evaluate retrofit potential in public school buildings. *Renewable and Sustainable Energy Reviews* **110**:266–277 DOI 10.1016/j.rser.2019.04.073.
- Cerquitelli T, Malnati G, Apiletti D. 2019. Exploiting scalable machine-learning distributed frameworks to forecast power consumption of buildings. *Energies* **12**(15):2933 DOI 10.3390/en12152933.
- Chen Y, Tan H. 2017. Short-term prediction of electric demand in building sector via hybrid support vector regression. *Applied Energy* **204**(5):1363–1374 DOI 10.1016/j.apenergy.2017.03.070.
- Ciulla G, D’Amico A, Brano VL, Traverso M. 2019. Application of optimized artificial intelligence algorithm to evaluate the heating energy demand of non-residential buildings at european level. *Energy* **176**(242):380–391 DOI 10.1016/j.energy.2019.03.168.
- D’Amico A, Ciulla G, Traverso M, Brano VL, Palumbo E. 2019. Artificial neural networks to assess energy and environmental performance of buildings: an Italian case study. *Journal of Cleaner Production* **239**:117993 DOI 10.1016/j.jclepro.2019.117993.
- Elmqvist N, Dragicevic P, Fekete J-D. 2008. Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics* **14**(6):1141–1148 DOI 10.1109/TVCG.2008.153.
- Gao W, Alsarraf J, Moayedi H, Shahsavar A, Nguyen H. 2019. Comprehensive preference learning and feature validity for designing energy-efficient residential buildings using machine learning paradigms. *Applied Soft Computing* **84**(3):105748 DOI 10.1016/j.asoc.2019.105748.
- Geyer P, Singaravel S. 2018. Component-based machine learning for performance prediction in building design. *Applied Energy* **228**:1439–1453 DOI 10.1016/j.apenergy.2018.07.011.
- Hastier T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. New York: Springer DOI 10.1007/978-0-387-84858-7.
- Heddam S. 2016. Multilayer perceptron neural network-based approach for modeling phycoyanin pigment concentrations: case study from lower charles river buoy, USA. *Environmental Science and Pollution Research* **23**(17):17210–17225 DOI 10.1007/s11356-016-6905-9.
- Himeur Y, Alsalemi A, Bensaali F, Amira A. 2020a. Building power consumption datasets: survey, taxonomy and future directions. *Energy and Buildings* **227**(6):page–110404 DOI 10.1016/j.enbuild.2020.110404.

- Himeur Y, Alsalemi A, Bensaali F, Amira A. 2020b.** A novel approach for detecting anomalous energy consumption based on micro-moments and deep neural networks. *Cognitive Computation* **12**(6):1381–1401 DOI [10.1007/s12559-020-09764-y](https://doi.org/10.1007/s12559-020-09764-y).
- Himeur Y, Alsalemi A, Bensaali F, Amira A. 2020c.** Robust event-based non-intrusive appliance recognition using multi-scale wavelet packet tree and ensemble bagging tree. *Applied Energy* **267**(1):114877 DOI [10.1016/j.apenergy.2020.114877](https://doi.org/10.1016/j.apenergy.2020.114877).
- IESVE. 2008.** Integrated environmental solutions virtual environment. Available at <https://www.iesve.com/> (accessed 9 December 2021).
- Kumar S, Pal SK, Singh RP. 2018.** Intra elm variants ensemble based model to predict energy performance in residential buildings. *Sustainable Energy, Grids and Networks* **16**(10):177–187 DOI [10.1016/j.segan.2018.07.001](https://doi.org/10.1016/j.segan.2018.07.001).
- Le LT, Nguyen H, Dou J, Zhou J. 2019a.** A comparative study of pso-ann, ga-ann, ica-ann, and abc-ann in estimating the heating load of buildings' energy efficiency for smart city planning. *Applied Sciences* **9**(13):2630 DOI [10.3390/app9132630](https://doi.org/10.3390/app9132630).
- Le LT, Nguyen H, Zhou J, Dou J, Moayedi H. 2019b.** Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique pso-xgboost. *Applied Sciences* **9**(13):2714 DOI [10.3390/app9132714](https://doi.org/10.3390/app9132714).
- Li X, Ying Y, Xu X, Wang Y, Hussain SA, Hong T, Wang W. 2020.** Identifying key determinants for building energy analysis from urban building datasets. *Building and Environment* **181**:107114 DOI [10.1016/j.buildenv.2020.107114](https://doi.org/10.1016/j.buildenv.2020.107114).
- Li Z, Dai J, Chen H, Lin B. 2019.** An ann-based fast building energy consumption prediction method for complex architectural form at the early design stage. *Building Simulation* **12**(4):665–681 DOI [10.1007/s12273-019-0538-0](https://doi.org/10.1007/s12273-019-0538-0).
- Moayedi H, Bui DT, Dounis A, Lyu Z, Foong LK. 2019.** Predicting heating load in energy-efficient buildings through machine learning techniques. *Applied Sciences* **9**(20):4338 DOI [10.3390/app9204338](https://doi.org/10.3390/app9204338).
- Naji S, Keivani A, Shamshirband S, Alengaram UJ, Jumaat MZ, Mansor Z, Lee M. 2016.** Estimating building energy consumption using extreme learning machine method. *Energy* **97**(5):506–516 DOI [10.1016/j.energy.2015.11.037](https://doi.org/10.1016/j.energy.2015.11.037).
- Navarro-Gonzalez FJ, Villacampa Y. 2019.** An octahedric regression model of energy efficiency on residential buildings. *Applied Sciences* **9**(22):4978 DOI [10.3390/app9224978](https://doi.org/10.3390/app9224978).
- Ngo N-T. 2019.** Early predicting cooling loads for energy-efficient design in office buildings by machine learning. *Energy and Buildings* **182**(1):264–273 DOI [10.1016/j.enbuild.2018.10.004](https://doi.org/10.1016/j.enbuild.2018.10.004).
- Nilashi M, Dalvi-Esfahani M, Ibrahim O, Bagherifard K, Mardani A, Zakuan N. 2017.** A soft computing method for the prediction of energy performance of residential buildings. *Measurement* **109**(3):268–280 DOI [10.1016/j.measurement.2017.05.048](https://doi.org/10.1016/j.measurement.2017.05.048).
- Pham A-D, Ngo N-T, Truong TTH, Huynh N-T, Truong N-S. 2020.** Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production* **260**(1):121082 DOI [10.1016/j.jclepro.2020.121082](https://doi.org/10.1016/j.jclepro.2020.121082).
- Roy SS, Samui P, Nagtode I, Jain H, Shivaramakrishnan V, Mohammadi-Ivatloo B. 2020.** Forecasting heating and cooling loads of buildings: a comparative performance analysis. *Journal of Ambient Intelligence and Humanized Computing* **11**(3):1253–1264 DOI [10.1007/s12652-019-01317-y](https://doi.org/10.1007/s12652-019-01317-y).
- Rumelhart DE, Hinton GE, Williams RJ. 1985.** Learning internal representations by error propagation. Technical report. California Univ San Diego La Jolla Inst. for Cognitive Science.

- Sadeghi A, Younes Sinaki R, Young WA, Weckman GR. 2020.** An intelligent model to predict energy performances of residential buildings based on deep neural networks. *Energies* **13**(3):571 DOI [10.3390/en13030571](https://doi.org/10.3390/en13030571).
- Seyedzadeh S, Rahimian FP, Rastogi P, Glesk I. 2019.** Tuning machine learning models for prediction of building energy loads. *Sustainable Cities and Society* **47**:101484 DOI [10.1016/j.scs.2019.101484](https://doi.org/10.1016/j.scs.2019.101484).
- Sharif SA, Hammad A. 2019.** Developing surrogate ann for selecting near-optimal building energy renovation methods considering energy consumption, LCC and ICA. *Journal of Building Engineering* **25**(November 2018):100790 DOI [10.1016/j.jobe.2019.100790](https://doi.org/10.1016/j.jobe.2019.100790).
- Tian W, Yang S, Zuo J, Li Z, Liu Y. 2017.** Relationship between built form and energy performance of office buildings in a severe cold Chinese region. *Building Simulation* **10**(1):11–24 DOI [10.1007/s12273-016-0314-3](https://doi.org/10.1007/s12273-016-0314-3).
- Tien Bui D, Moayedi H, Anastasios D, Kok Foong L. 2019.** Predicting heating and cooling loads in energy-efficient buildings using two hybrid intelligent models. *Applied Sciences* **9**(17):3543 DOI [10.3390/app9173543](https://doi.org/10.3390/app9173543).
- Tsanas A, Xifara A. 2012.** Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings* **49**(3):560–567 DOI [10.1016/j.enbuild.2012.03.003](https://doi.org/10.1016/j.enbuild.2012.03.003).
- Uyank GK, Güler N. 2013.** A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences* **106**:234–240 DOI [10.1016/j.sbspro.2013.12.027](https://doi.org/10.1016/j.sbspro.2013.12.027).
- Wang N, Phelan PE, Gonzalez J, Harris C, Henze GP, Hutchinson R, Langevin J, Lazarus MA, Nelson B, Pyke C, Roth K, Rouse D, Sawyer K, Selkowitz S. 2017.** Ten questions concerning future buildings beyond zero energy and carbon neutrality. *Building and Environment* **119**:169–182 DOI [10.1016/j.buildenv.2017.04.006](https://doi.org/10.1016/j.buildenv.2017.04.006).
- Xu C, Chen H. 2020.** A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data. *Energy and Buildings* **215**(9):109864 DOI [10.1016/j.enbuild.2020.109864](https://doi.org/10.1016/j.enbuild.2020.109864).
- Yeom S, Kim H, Hong T, Lee M. 2020.** Determining the optimal window size of office buildings considering the workers' task performance and the building's energy consumption. *Building and Environment* **177**(3):106872 DOI [10.1016/j.buildenv.2020.106872](https://doi.org/10.1016/j.buildenv.2020.106872).
- Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y, Livingood W. 2021.** A review of machine learning in building load prediction. *Applied Energy* **285**(6245):116452 DOI [10.1016/j.apenergy.2021.116452](https://doi.org/10.1016/j.apenergy.2021.116452).
- Zhu S, Heddam S, Wu S, Dai J, Jia B. 2019.** Extreme learning machine-based prediction of daily water temperature for rivers. *Environmental Earth Sciences* **78**(6):1–17 DOI [10.1007/s12665-019-8202-7](https://doi.org/10.1007/s12665-019-8202-7).