A machine-learning surrogate model for optimising photovoltaic-thermal deployment in complex urban morphologies

eceee Summer Study proceedings

COLLECTION: BUILDINGS AND APPLIANCES

PEER-REVIEWED PAPER

ALIREZA NAZERI

CIARA AHERN

*Author affiliations can be found in the back matter of this article

ABSTRACT

Peer-reviewed paper 8-244-26

Photovoltaic-thermal (PVT) systems can simultaneously generate electricity and recover usable heat, significantly improving the utilisation of limited urban roof space. However, their performance in dense cities is strongly influenced by urban morphology and installation geometry, factors not adequately captured by conventional photovoltaic forecasting approaches. This study proposes a simulation-trained machine learning (ML) framework designed to optimise PVT layouts by evaluating key parameters, including panel density (inter-panel spacing), angles of inclination, and orientation, to achieve optimal electrical and thermal generation across varying urban environments.

A parametric dataset of 1,575 installation scenarios was generated using GIS-based urban modelling and building performance simulations across three morphologically distinct districts of Tehran. Artificial Neural Network (ANN) and Random Forest (RF) models were trained to predict electricity generation, hot water production, and associated carbon reduction (CR). Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) metrics.

The RF model achieved high predictive accuracy ($R^2 = 0.91–0.99$) with shorter training times than the ANN. Sensitivity analysis revealed that urban form and installation geometry (panel spacing and tilt) exert a stronger influence on combined energy yield than several climatic predictors. By replacing computationally intensive simulations with a data-driven surrogate, the proposed framework enables rapid design-space exploration and provides a robust tool for informed PVT deployment in energy-constrained urban environments.

CORRESPONDING AUTHOR:

Alireza Nazeri

School of Mechanical Engineering, Technological University Dublin, Dublin, Ireland; Irish Building Stock Observatory, Dublin, Ireland

alirezanazeri95@gmail.com

KEYWORDS:

photovoltaic-thermal (PVT) systems; machine learning in renewable energy; energy efficiency optimisation; AI-driven energy prediction

TO CITE THIS ARTICLE:

Nazeri, A and Ahern, C. 2026. A machine-learning surrogate model for optimising photovoltaic-thermal deployment in complex urban morphologies. *eceee Summer Study proceedings*, 1(1): 29, pp. 1–22. DOI: https://doi.org/10.66506/essp.8-244-26

INTRODUCTION

Optimising the installation patterns of photovoltaic (PV) systems on building envelopes is crucial for maximising energy generation efficiency and system performance. Recent research highlights the significance of factors such as installation orientation, tilt angle Optimisation, and mitigation of shading effects. Studies emphasise that selecting the appropriate tilt angle based on geographical location can significantly enhance solar capture by minimising angle-related losses (Jing et al. 2023). Furthermore, strategies to minimise shading effects are essential for maintaining consistent energy production throughout the day (Wang et al. 2021). Integrating PV systems into building facades or roofs not only optimizes energy generation but also contributes to sustainable urban development, aligning architectural and aesthetic considerations with renewable energy goals (Chiang et al. 2022).

Recent advancements in artificial intelligence (AI) and machine learning (ML) have significantly advanced the optimisation and prediction capabilities of photovoltaic-thermal (PVT) systems in urban environments (Dev et al. 2022). These advanced techniques enable better optimisation and adaptation of PVT systems, particularly in dynamic urban environments, highlighting their growing relevance in enhancing renewable energy technologies.

This study employs AI and ML techniques, specifically ANN and Random Forest (RF) algorithms, to develop a predictive framework for PVT systems within complex urban contexts. This research integrates predictive models that simultaneously evaluate electrical output, thermal performance, and environmental impact. By incorporating the carbon reduction (CR) potential of these systems in dense settings, the study offers a holistic assessment of their operational sustainability. To support this objective, a modelling framework was devised using representative urban building blocks of varying heights to analyse energy demand and solar absorption potential. The primary aim is not merely to observe performance trends, but to validate a data-driven surrogate model that can function as a decision-support tool. This platform transforms complex simulation outputs into actionable recommendations, lowering the barrier to integrating hybrid solar systems by providing rapid, evidence-based insights for non-expert stakeholders. Specifically, the study culminates in a four-step decision-support framework (Figure 15) that integrates GIS-based urban modelling with a trained RF surrogate model, enabling non-expert users to input site-specific parameters and receive near-instantaneous predictions of EP, HWP, and CR without the need for computationally intensive physical simulations.

The predictive capability of the models is evaluated using established statistical indicators. To ensure the framework's reliability, the study addresses the following research questions:

1. Can a data-driven model accurately replicate the non-linear performance variations of PVT systems within complex urban morphologies?
2. To what extent can a trained algorithm (RF or ANN) effectively replace computationally intensive physical simulations for non-expert users?

The rationality of the sensitivity analysis presented herein serves as a fundamental validation of the framework. By demonstrating that the model has correctly "learned" the physical interrelationships between installation geometry and energy output, the study establishes the model's reliability. The novelty of this work, therefore, lies in the creation of a validated, replicable methodology that democratises access to sophisticated solar-thermal forecasting in dense urban environments.

LITERATURE REVIEW

This literature review examines the application of various Machine Learning (ML) algorithms for predicting and optimising solar power generation across building, urban, and industrial scales as detailed in the provided sources (Table 1). Using ML has several benefits compared to traditional methods, such as: Superior Accuracy under Variable Conditions, Design-Space Optimisation, Integrated Multi-Output Prediction, Handling Complex Urban Interactions and Rapid Evaluation and Optimisation (Pham and Tran 2023).

REFERENCE	CONTEXT	OUTPUT INDICATORS	ALGORITHM OF ML	MODULE TYPE	MODELLING SOFTWARE	VALIDATION STRATEGY	ACCURACY METRIC	BEST ACCURACY	DATA TYPE	VARIABLE
Shin et al. (2022)	Building	power generation	ANN	BIPV	PVsyst, Solar Pro	Train/test	RMSE, MAE, R^2	0.92 (R^2)	Simulated, Measured	Solar irradiance
Serrano-Luján et al. (2022)	Building	power generation	GE, DE	BIPV	_	Train/test	Erel	1.55	Measured	Ambient Temperature, solar irradiation, relative outdoor humidity, wind speed
Yousif and Kazem (2021)	Building	power generation	ANN	PVT	_	Train/test	MSE, NMSE, MAE, R^2	0.11 (NMSE)	Measured	solar irradiance, ambient temperature
Sulaiman et al. (2024)	Building	total active power	NN	PV	_	Train/test	RMSE, MAE, Standard deviation	0.7 RMSE	Measured	Ambient Temperature, Horizontal irradiation
Mohana et al. (2021)	Building	Power generation	LASSO, RF, LR, PR, XGBoost, SVM, NN	PV	_	k-fold cross-validation	MSE	0.9	Measured	External temperature, wind speed, Humidity
Chaibi et al. (2021)	Building	electrical and thermal efficiencies	ANN	PVT	_	Train/test	MAE	0.0078%	Measured	solar irradiance and the module temperature
Patel et al. (2022)	Urban	power generation	LR	PV	_	Train/test	MAPE	1.40%	Measured	Ambient temperature, Relative Humidity
Kassem and Othman (2022)	Building	power generation	MFFNN, CFNN, RBFNN, ENN	PV	Matlab	Train/test	R, MAE, RMSE	0.0021 (RMSE)	Dataset	Ambient temperature, relative Humidity, solar radiation, wind speed
Suanpang and Jamjuntr (2024)	Microgrid	power generation	LGBM, KNN	PV	_	Train/test	R^2, RMSE, MAE	0.84 R^2	Measured	Solar irradiance, ambient temperature
Cao et al. (2022)	Building	electrical efficiency	ANFIS, ANN, LS-SVR	PVT	_	Train/test	AARD, MSE, R^2	0.95 R^2	Dataset	radiation intensity, Coolant material
Zazoum (2022)	Module	power generation	SVM, GPR	PV	_	Train/test	RMSE, MAE, R^2	0.98 R^2	Dataset	Panel temperature, ambient temperature, relative humidity,
Zhang et al. (2021)	Module	power generation	NN	PV	_	Train/test	MSE	0.01	Measured	Radiation, Ambient temperature, Humidity, Wind speed, Evaporation
Pham and Tran (2023)	Building	power generation	KNR, LASSO, SVR, RF, ETR, GBR, XGBoost, ANN	PV	_	Train/test	MAE, RMSE	0.60 (RMSE)	Dataset	Radiation, Ambient temperature, Humidity, Wind speed, Evaporation, Rooftop dimension
Gharaee et al. (2024)	Module	electrical efficiency	MLP, RF, SVR	PVT	_	Train/test	RMSE, R^2	0.76 (R^2)	Measured	mass flow rate, solar radiation, ambient temperature, wind speed, fluid inlet temperature, PVT surface area, pipe inner diameter
Zhou et al. (2020)	power plant	power generation	ELM, GA, SDA	PV	_	Train/test	R^2, MAE, nRMSE	0.59 (R^2)	Dataset	daily maximal, minimal and averaged temperature, daily averaged global horizontal radiation, daily averaged diffusive horizontal radiation

(Contd.)

REFERENCE	CONTEXT	OUTPUT INDICATORS	ALGORITHM OF ML	MODULE TYPE	MODELLING SOFTWARE	VALIDATION STRATEGY	ACCURACY METRIC	BEST ACCURACY	DATA TYPE	VARIABLE
Rojek et al. (2023)	Building	Power generation, CO_2 reduction	ANN	PV	_	Train/test	RMSE	0.01	Dataset	air temperature, wind speed, cloudiness, Current power direction
Alghamdi et al. (2023)	Module	power generation	DNN	PVT	_	Train/test	MSE	3.34E-08	Measured	Cell type
Scott et al. (2023)	Building	power generation	RF, NN, SVM, LR	PV	_	Train/test	RMSE, MAPE	1.76 (RMSE)	Measured	weather and solar generation data
Tripathi et al. (2024)	power plant	power generation	MR, SVMR, GR	PV	_	Train/test	MSE, MAE, R^2	0.88 (R^2)	Measured	solar radiation, ambient temperature, relative humidity
Kabilan et al. (2021)	Building	power generation	ANN, DT, QSVM	BIPV	_	Train/test	RMSE, MSE, R^2, MAPE, MAE	0.88 (R^2)	Measured	Building orientations
Elsaraiti and Merabet (2022)	power plant	power generation	LSTM, MLP	PV	_	Train/test	MAE, MAPE, RMSE, R^2	0.77 (R^2)	Dataset	weather and solar generation data
Asiedu et al. (2024)	power plant	power generation	ANN, XGBoost, RF, DT, KNN, LASSO, LR, RR	PV	_	Train/test	RMSE, MAE, R^2	0.84 (R^2)	Dataset	ambient temperature, module temperature, irradiation

Table 1 Related research.

DOI: 10.66506/essp.8-244-26

BUILDING-INTEGRATED AND URBAN POWER GENERATION

Several studies focus on predicting power generation within the built environment, using Building-Integrated Photovoltaics (BIPV) and standard rooftop systems. Shin et al. (2022) utilised an ANN combined with PVsyst® and Solar® Pro software to model BIPV generation, achieving a high accuracy with an R^2 of 0.92. Similarly, Serrano-Luján et al. (2022) applied Genetic Algorithms (GA) and Differential Evolution (DE) for BIPV power prediction, using measured ambient temperature and solar irradiation as key variables.

For general building power generation, Sulaiman et al. (2024) achieved an RMSE of 0.7 using a Neural Network (NN), while Mohana et al. (2021) compared multiple models including LASSO, RF, and XGBoost, reaching an MSE of 0.9. Kassem and Othman (2022) explored specialised architectures like MFFNN and RBFNN in Matlab, attaining an impressive RMSE of 0.0021. In the context of urban power generation, Patel et al. (2022) used Linear Regression to achieve a MAPE of 1.40%. Other research by Pham and Tran (2023) incorporated structural data like rooftop dimensions alongside weather variables, testing various models such as KNR and GBR to reach an RMSE of 0.60.

PV AND PVT EFFICIENCIES

A subset of the literature emphasises the thermal and electrical efficiency of specialised modules. Yousif and Kazem (2021) applied an ANN to PVT systems with a best accuracy of 0.11 NMSE. Chaibi et al. (2021) also utilised ANNs to model building electrical and thermal efficiencies, reporting a very low MAE of 0.0078%. Efficiency modelling was further advanced by Cao et al. 2022 who used ANFIS and LS-SVR models to achieve an R^2 of 0.95, focusing on radiation intensity and coolant materials as variables. More recently, Gharaee et al. (2024) studied PVT module efficiency using MLP and RF, identifying mass flow rate and fluid inlet temperature as critical inputs to achieve an R^2 of 0.76.

MODULE-LEVEL POWER AND PERFORMANCE

Predicting the specific output of individual modules is a frequent theme. Zazoum (2022) employed SVM and Gaussian Process Regression (GPR) to model module power generation, reaching an R^2 of 0.98 based on panel and ambient temperatures. Zhang et al. (2021) focused on the impact of evaporation and wind speed using a Neural Network, yielding an MSE of 0.01. Alghamdi et al. (2023) achieved the highest precision recorded in the sources (MSE of 3.34E-08) using a Deep Neural Network (DNN) to model module power based on cell type.

Large-scale power plant and microgrid applications

For larger installations, Suanpang and Jamjuntr (2024) modelled microgrid power generation using LGBM and KNN, achieving an R^2 of 0.84. Regarding power plants:

- Zhou et al. (2020) utilised ELM, GA, and SDA algorithms, achieving an R^2 of 0.59.
- Tripathi et al. (2024) applied Support Vector Machine Regression (SVMR) and reached an R^2 of 0.88.
- Elsaraiti and Merabet (2022) focused on time-series forecasting using LSTM and MLP, reporting an R^2 of 0.77.
- Asiedu et al. (2024) conducted a comprehensive comparison of eight different models, including XGBoost and Ridge Regression, resulting in an R^2 of 0.84.

METHODOLOGICAL TRENDS

Across these studies, ANN and RF are among the most frequently cited algorithms for their robustness in handling non-linear weather data. The most common validation strategy is the train/test split, though k-fold cross-validation was notably used by Mohana et al. (2021). The primary variables driving these models consistently include solar irradiance (or radiation), ambient temperature, and humidity. Additionally, Rojek et al. (2023) demonstrated the utility of ML in predicting not only power but also CO_2 reduction, highlighting the environmental impact of these technological applications.

DOI: 10.66506/essp.8-244-26

RESEARCH GAPS

Several limitations in previous research (Table 1) identified regarding solar power prediction:

- Focus on Single Parameters: Prior research often focuses on a single indicator, typically exclusively on PV electricity production (EP), while neglecting other vital outputs such as environmental impacts.
- Reliance on Historical Performance Data: Most existing studies rely on datasets of actual, historical PV panel performance rather than predictive modelling for specialised systems.
- Neglect of Thermal Outputs: Despite the dual nature of PVT systems, many studies fail to integrate predictive models for both electrical and thermal outputs simultaneously.
- Climate-Centric Variables: Much of the existing literature uses climatic indicators (such as humidity, wind speed, or ambient temperature) as the primary variables, often overlooking the impact of geometric installation conditions and urban morphology.
- Limited Environmental Integration: There is a lack of comprehensive analysis that combines energy yield with integrated environmental impacts, such as CR, in an urban context.

RESEARCH CONTRIBUTIONS

This study introduces a multi-faceted approach to address these deficiencies:

- Integrated multi-output modelling: Unlike previous studies, this research differentiates itself by integrating predictive models for three distinct indicators: electricity production (EP), hot water production (HWP), and CR.
- Focus on Installation Patterns: Instead of focusing solely on weather data, this study examines the effect of 1,575 different installation configurations, including horizontal/vertical angles, panel-to-panel distance, and elevation from the roof.
- Impact of Urban Morphology: The study models these configurations across three unique urban districts in Tehran, which has cold semi-arid climate, with varying building heights and densities (from low-rise/high-density to high-rise/low-density) to account for shading effects in a dynamic urban environment.
- AI-Driven Optimisation: By leveraging and comparing ANN and RF, the study identifies that the RF algorithm provides superior performance for this integrated modelling, achieving an R^2 accuracy of 0.91 and significantly faster learning times.
- Development of a Non-Expert Framework: The research culminates in a four-step software framework that integrates GIS data and AI. This tool is designed to facilitate decision-making for non-experts, allowing them to optimise PVT installations based on specific urban areas and panel properties.
- Comprehensive Environmental Assessment: The study explicitly calculates the environmental benefits by correlating electricity and hot water production (HWP) with specific carbon emission reduction metrics.

METHODOLOGY

The study followed a systematic procedure: First, a dataset based on the variables was generated using simulation-based software. This dataset contains all the possibilities considered for installing PVT on rooftop (Table 2). In next step, ML algorithms were applied to analyse this dataset and develop a predictive model. Subsequently, a new set of variables, not present in the original dataset, was introduced to assess the model's performance with novel input data. This phase involved an iterative trial-and-error process to optimise the performance of the ML algorithms. Finally, a sensitivity analysis was performed to identify which features most significantly affected the calculated metrics (Figure 1).

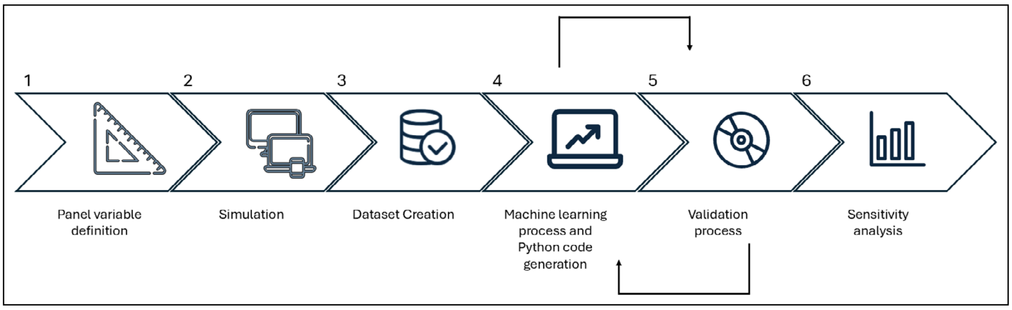

Figure 1 Schematic diagram of research steps.

CONTEXT

To create the dataset, Tehran, was chosen which is divided into 22 distinct urban areas (Figure 2), each with unique characteristics. Three selected regions are central to this research, with the selection criterion focusing on variations in elevation and population density across the city.

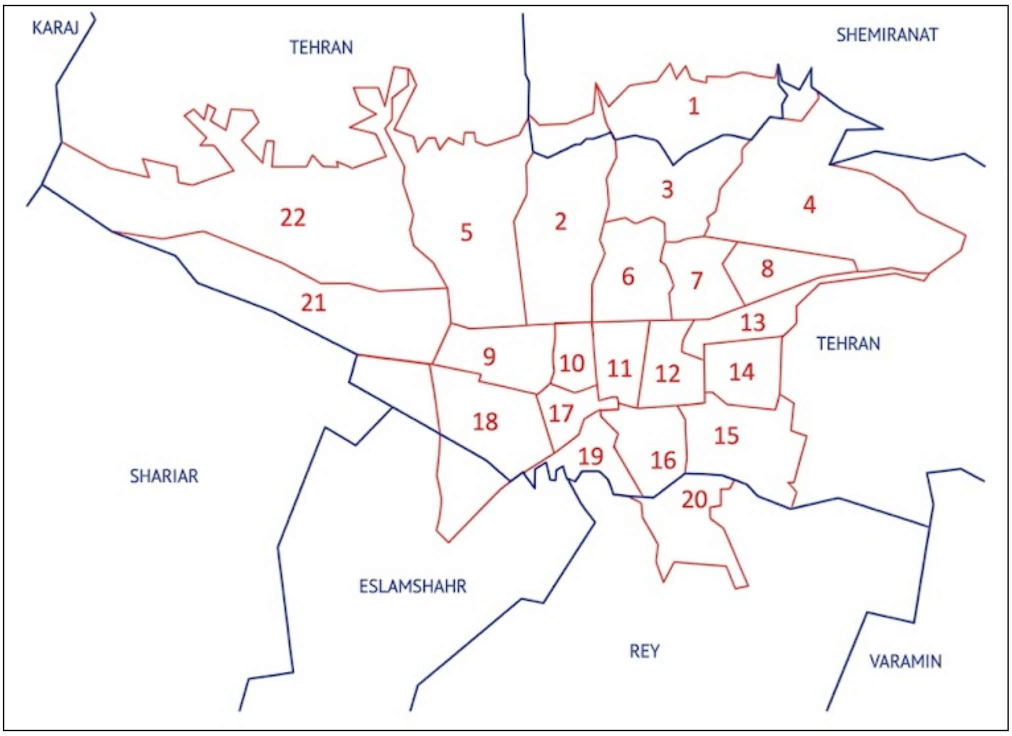

Figure 2 Map of Tehran's urban divisions.

Three morphologically distinct regions were selected: Region 1 (high-rise/dense) Figure 3, Region 2 (mid-rise/dispersed) Figure 4, and Region 16 (low-rise/dense packed) Figure 5.

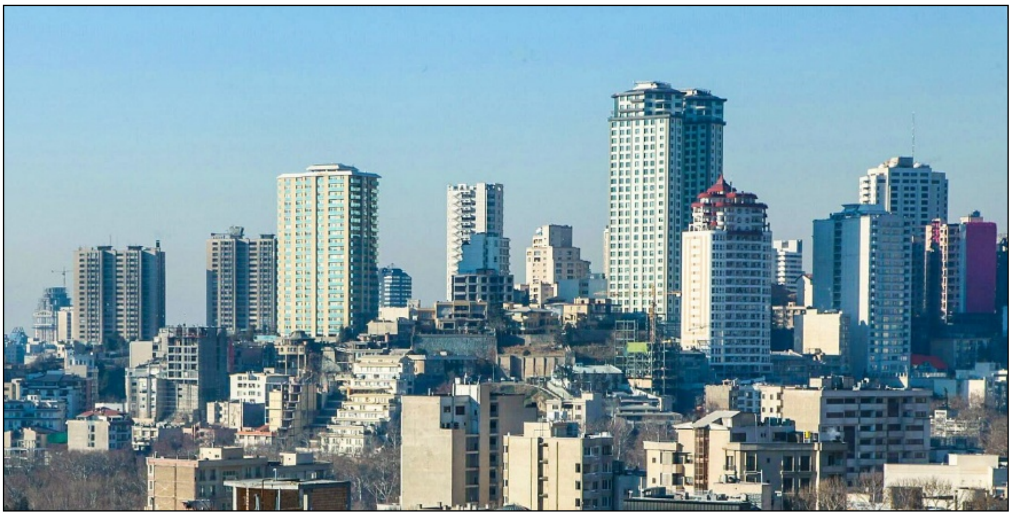

Figure 3 Urban morphology of region 1.

eceee Summer Study proceedings
DOI: 10.66506/essp.8-244-26

Figure 4 Urban morphology of region 2.

Figure 5 Urban morphology of region 16.

DATASET CREATION AND MODELLING

To standardise calculations and ensure comparability of results, a specific type of thermal photovoltaic panel was utilised. The chosen model is the PV-MLE275HD2, manufactured by Mitsubishi. This panel is equipped with monocrystalline silicon photovoltaic cells, with 120 cells per panel. With a power output of 275 watts and a maximum voltage supply of 32.1 volts, these cells demonstrate an overall efficiency rating of 16.6%.

To construct a robust dataset for the ML surrogate model, 1,575 unique installation configurations were systematically generated (Table 2). These configurations represent the «potential states» of the system, defined by five primary variable parameters that directly influence both electrical and thermal performance within dense urban environments.

DOI: 10.66506/essp.8-244-26

- Geometric Variables: The Horizontal Angle and Vertical Angle were adjusted to optimise solar azimuth and tilt, thereby minimising angular losses relative to specific geographic position. Furthermore, Panel Elevation was incorporated to account for convective cooling, a critical factor in maintaining PV cell efficiency by preventing overheating.
- Spatial Configuration: The Distance Between Panels serves as a vital parameter for assessing self-shading effects within the array. This variable allows the model to learn the critical trade-off between maximising panel density on a limited roof area and minimising inter-row shading losses.
- Urban Morphology: To ensure the model remains accurate across diverse cityscapes, these configurations were modelled across three distinct urban districts.

Table 2 Selected parameters and their values.

VARIABLE PARAMETERS	RANGE OF VARIATION	STEP OF VARIATION	NUMBER OF CASES	DESCRIPTIONS
Horizontal Angle of panel	–20° to 20°	10°	5	
Vertical angle of panel	15° above and below the latitude	5°	7	
Distance Between Panels	1 to 5 meters	1 meter	5	
Panel elevation from roof	10 to 30 centimetres	10 centimetres	3	
Urban Block	–	–	3	In three different regions
Dataset variables number	5 × 7 × 5 × 3 × 3 = 1,575			

To initiate modelling of urban maps GIS files for regions were obtained from the Tehran Municipality. Utilising ArcGIS software, a shapefile file was generated for the designated urban blocks. The shapefile format serves as a straightforward and indirect method for storing geometric and geographic feature information.

In next step for solar radiation analysis, the Ladybug plugin within the Grasshopper software was utilised. This plugin, powered by the EnergyPlus engine, was employed to perform PVT output calculations. The solar radiation analysis algorithm was designed using Tehran's epw file for 2020, providing essential meteorological data. Then, surfaces with an average annual radiation exceeding 1,500 kWh/m^2/yr were identified and compiled into a list. This threshold is derived from the findings of Biyik et al. (2017). Surfaces with radiation below 1,500 kWh/m^2/yr, wouldn't be efficient for installing PV cells. Then the focus shifts to the installation of thermal photovoltaic panels on surfaces identified as optimal for radiation reception in previous stages. To facilitate this process, an algorithm was developed within the Grasshopper environment.

Panel geometry was parametrically varied across all 1,575 configurations using the Colibri plugin in Grasshopper, which iteratively evaluated all parameter combinations to generate the thermal, electrical, and environmental outputs. Upon specifying the desired parameters and installing the panels, the thermal and electrical outputs of each panel are calculated. Additionally, environmental outputs are generated to provide a comprehensive assessment of the system›s performance and sustainability.

For the environmental impact assessment, a life-cycle approach was adopted to quantify the net CR potential of the PVT system. This calculation accounts for the carbon emissions offset by both the electrical and thermal yields of the panels. In the context of Tehran, where residential water heating is almost exclusively powered by natural gas, the thermal energy produced by the PVT panels directly offsets the domestic gas load. It is assumed that each unit of thermal energy generated by the system displaces an equivalent amount of fossil fuel consumption. Consequently, the CR is calculated by estimating the volume of natural gas that would otherwise be burned to meet this heating demand, thereby determining the resultant greenhouse gas (GHG) emissions avoided on an annual basis. Furthermore, the electricity generated by the panels contributes to carbon mitigation by displacing grid-supplied electricity, which carries a specific carbon intensity per kilowatt-hour (kWh) based on the local power plant mix. To ensure a balanced assessment, these gains are weighed against the 'carbon debt' of the system; specifically, the embodied carbon emissions associated with the manufacturing and production of each PVT panel. By determining the total number of panels utilised in each urban configuration, the net carbon benefit is calculated by subtracting the production-related emissions from the total annual avoided emissions (displaced gas and electricity). To quantify the net annual CR, the following equation was applied:

$$CR = (EP \times \alpha) + (HWP \times \beta) - (N \times \gamma) \quad (1)$$

Where EP is annual electricity production (kWh), HWP is annual hot water production (kWh), N is the number of panels, α is the grid electricity carbon intensity (0.183 $kgCO_2$/kWh), β is the natural gas displacement emission factor (0.185 $kgCO_2$/kWh thermal), and γ is the embodied carbon per panel ($kgCO_2$/panel). The value of α is consistent with Iran's national grid emission factor, while β reflects the CO_2 equivalent of natural gas combustion for domestic water heating. CR is expressed in kg CO_2 per year. Figure 6 illustrates the steps of creating dataset.

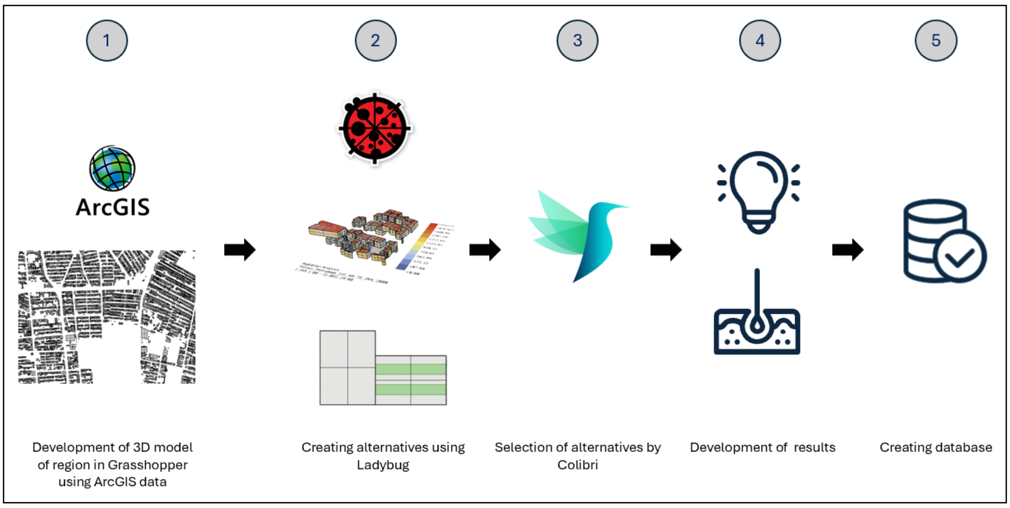

Figure 6 Steps for creating dataset.

ML ALGORITHM DEVELOPMENT

Following the simulation phase, a comprehensive dataset comprising 1,575 distinct scenarios was generated based on the parametric variables defined in Table 2. The resulting PVT performance data were curated and formatted to serve as the foundational input for the ML stage. In this research, the employed ML framework utilised supervised learning and regression analysis, specifically invoking ANN and RF algorithms.

These two algorithms were selected to develop a high-speed "surrogate model" capable of providing instantaneous performance predictions. The objective was to create a data-driven tool that could bypass the need for computationally intensive physical simulations when evaluating new urban installation configurations. Each algorithm is characterised by its unique set of hyperparameters, the external configuration settings that govern the model's structure and the efficiency of the learning process. For the ANN, this involved tuning variables such as the number of hidden layers and neurons to capture complex non-linear relationships. For the RF, hyperparameters included the number of decision trees and the maximum depth of each tree, which are essential for ensuring model stability and preventing overfitting.

Initially, a specific range was assigned to each hyperparameter; through a systematic optimisation process, the configurations that yielded the highest predictive accuracy were identified (Table 3).

Table 3 Type of algorithms and hyperparameters.

TYPE OF ALGORITHM	HYPERPARAMETER	NUMBER OF SCENARIOS
ANN	Number of Hidden Layers	1,2,3
	Number of Neurons per Layer	1 to 200 in steps of 10
	Activation Function	RELU
RF	Number of Decision Trees	1 to 500
	Max_depth	1 to 10 and None

DOI: 10.66506/essp.8-244-26

To ensure the computational efficiency and stability of the training process, the input parameters were normalised using the Min-Max scaling method. This ensured that all parameter values were scaled to a range between 0 and 1, which is instrumental in accelerating convergence and improving the accuracy of the gradient-based learning algorithms. The formula utilised for this normalisation is defined as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

To develop and verify the predictive models, the dataset was systematically partitioned into distinct subsets for training and testing. Specifically, a test set comprising 20% of the total data was initially held out; this data was not utilised during any part of the training or hyperparameter optimisation phases, serving as an "unseen" benchmark to check for signs of overfitting and to reduce the likelihood of evaluation bias.

The remaining 80% of the data was then further segmented: approximately 90% was allocated for model training, while 10% was reserved for internal validation during the learning process. This stratified approach ensures the model is exposed to a vast majority of the simulated scenarios while maintaining rigorous subsets for ongoing performance assessment. By keeping the distribution of these sets consistent, the framework ensures that the final accuracy metrics are representative of the model's true generalisation capability across diverse urban configurations. To ensure full reproducibility, a fixed random seed of 10 was applied to all data partitioning operations. No stratification was applied during splitting, as the target variables are continuous and the dataset was generated through systematic parametric sampling, ensuring an inherently balanced distribution of input configurations across all subsets.

Accuracy measurement

Accuracy measurement indices are vital for evaluating the performance and training of ML models. These indices compare the predicted values with the values obtained in the simulation. Common accuracy measurement indices include RMSE, percentage error (PE), recognition coefficient index (R^2), and MSE (Willmott and Matsuura 2005). In this research, three error measurement indicators, RMSE, R^2, and MAE are utilised. The calculation formulas for each are as follows.

SENSITIVITY ANALYSIS

Statistics involves comprehending the interrelationships among variables within a dataset and determining the extent to which these relationships are influenced by one another. In the context of this study, a sensitivity analysis was conducted on the dataset of 1,575 simulated scenarios to evaluate the impact of installation parameters, specifically inter-panel distance, horizontal/vertical angles, and panel elevation, on the target outputs: EP, thermal energy HWP, and CR.

By examining the performance trends within the urban blocks of Tehran, it was observed that the data for these specific indicators changes linearly relative to the input features. Consequently, correlation calculation methods tailored for linear data with consistent performance are applicable. The most prevalent methods in this domain are the Pearson and Spearman coefficients (Gonçalves et al. 2020). In this research, the Pearson correlation coefficient was specifically employed to quantify the sensitivity of the performance indicators to each geometric and contextual variable. Pearson correlation was selected because inspection of the simulation outputs confirmed that the target indicators (EP, HWP, and CR) vary approximately linearly with respect to the input parameters within the defined ranges of this study, satisfying the key assumption of the Pearson method and making it an appropriate and interpretable choice for this dataset. This allowed for the identification of which parameters, such as shading-induced distance or tilt angle, exerted the most significant influence on the system's dual-output efficiency.

ACCURACY MEASUREMENT

Accuracy measurement indices are vital for evaluating the performance and training of machine learning models. These indices compare the predicted values with the values obtained in the simulation. Common accuracy measurement indices include root mean square error (RMSE), percentage error (PE), recognition coefficient index (R^2), and mean square error (MSE) (Willmott and Matsuura 2005).

In this research, three error measurement indicators, RMSE, R^2, and MAE are utilised. The calculation formulas for each are as follows:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2} \tag{3}$$

Recognition Coefficient Index (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{4}$$

Mean Absolute Error (MAE):

$$MAE = \frac{\Sigma\left|y_i - \hat{y}_i\right|}{n} \tag{5}$$

RESULTS

The characteristics of the output data for each considered indicator, encompassing both energy and environmental metrics, have been investigated.

ELECTRICITY PRODUCTION SIMULATION RESULTS

The highest numerical amount of EP was recorded in region 16 with low-rise buildings (Figure 7), while the lowest was observed in region 1 which includes high-rise buildings. In terms of average annual EP, panels in region 16 outperformed those in region 2 by approximately 38% and exceeded the production in region 1 by about 52%. These differences are directly attributable to inter-building shading, with Region 16's lower-rise morphology allowing significantly greater solar access to rooftop surfaces. Figures 7 and 8, show simulation results of EP for training and test data in 3 different regions respectively.

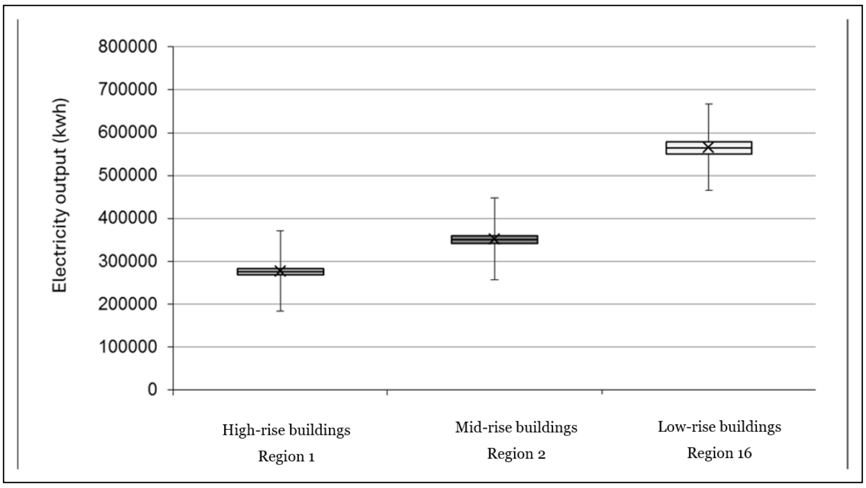

Figure 7 Simulation results of the EP index for training data in different regions.

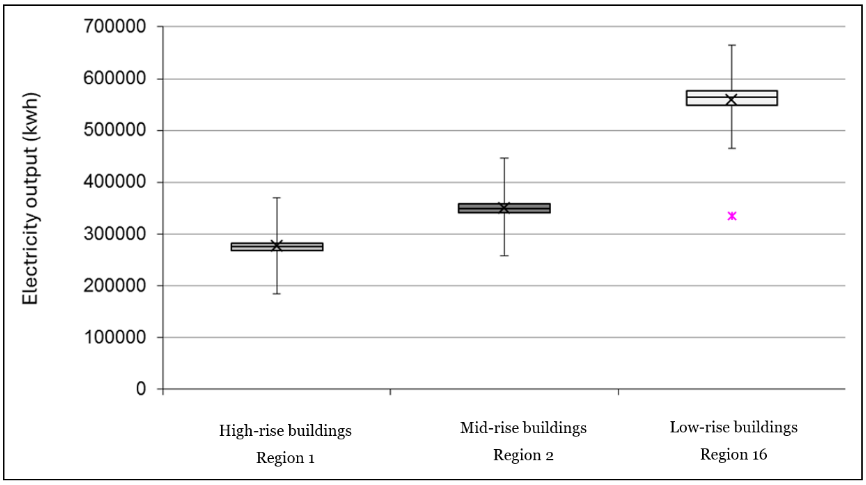

Figure 8 Simulation results of the EP index for test data in different regions.

HOT WATER PRODUCTION SIMULATION RESULTS

Region 2 recorded the highest numerical HWP (Figures 9 and 10), while Region 1 exhibited the lowest. In terms of average annual HWP, the performance of panels in region 16 surpasses that of region 2 by approximately 36% and exceeds that of region 1 by about 50%. The same shading mechanism governs thermal output, though the effect is less pronounced than for electricity production, as the thermal collector is less sensitive to partial shading than the PV cells. Furthermore, in regions characterised by more uniform building heights, (Region 16), HWP results are relatively consistent. In contrast, regions with varied building heights show significant differences in HWP, underscoring the impact of building height variability on system performance.

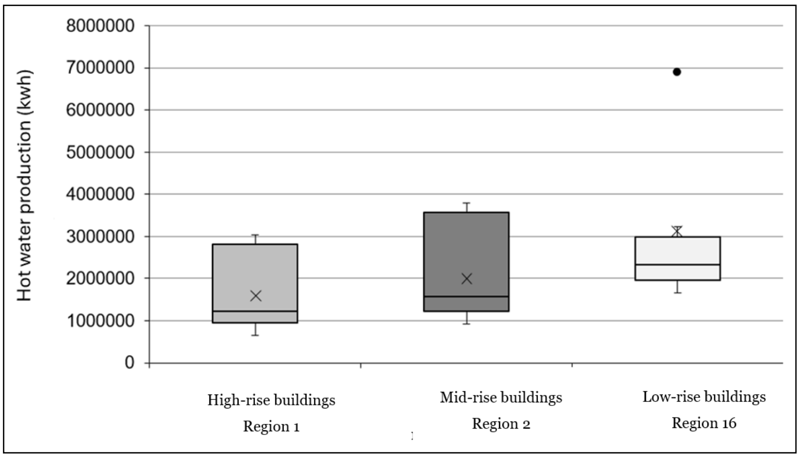

Figure 9 Simulation results of the HWP for training data in different regions.

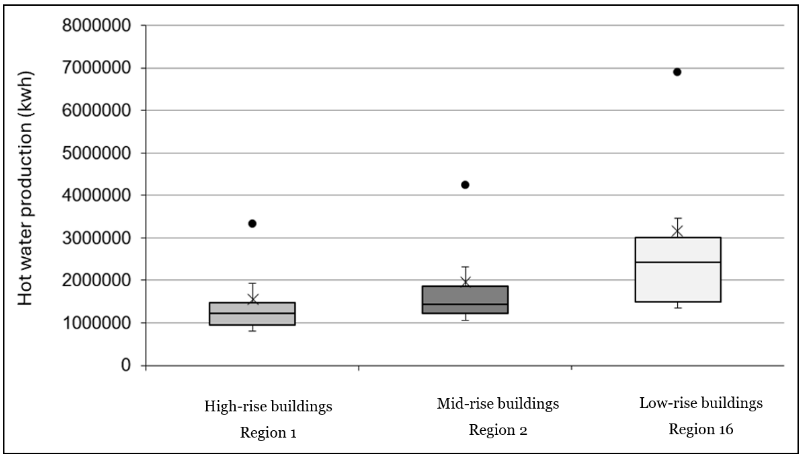

Figure 10 Simulation results of the HWP for test data in different regions.

CARBON REDUCTION SIMULATION RESULTS

On average, the highest amount of CR occurred in region 16 while region 1 exhibited the lowest average CR rate. Numerically, the highest amount of CR was observed in region 2, while the lowest was in region 1. Region 16 achieved a 56% higher CR compared to region 2, and approximately 90% higher compared to region 1 (Figures 11 and 12).

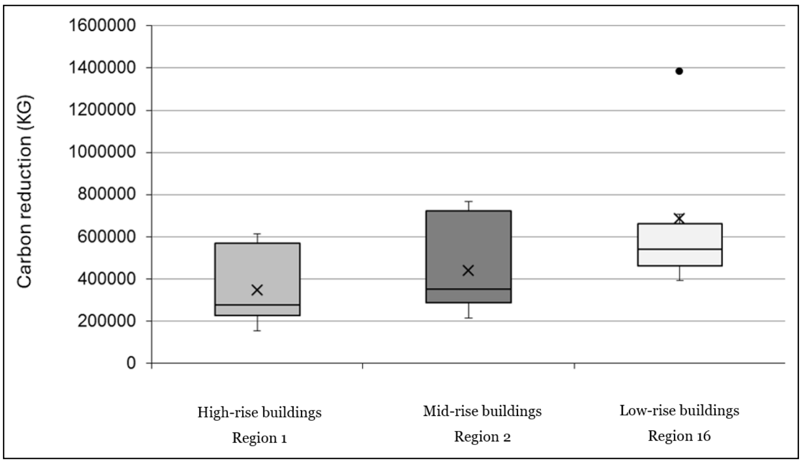

Figure 11 Simulation results of the CR for training data in different regions.

DOI: 10.66506/essp.8-244-26

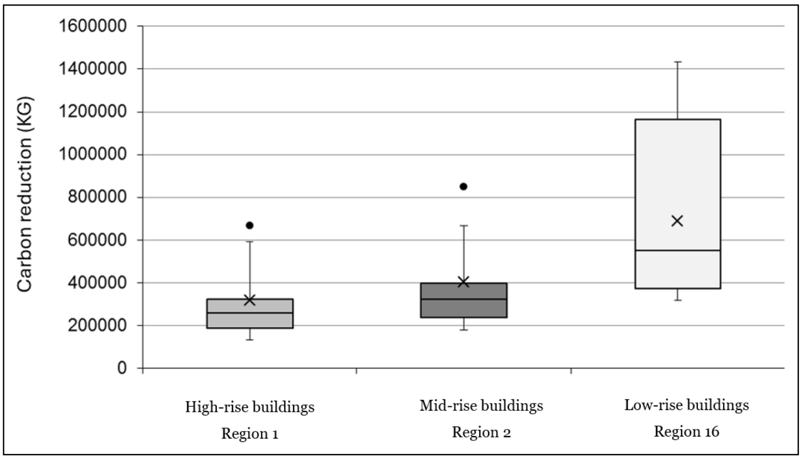

Figure 12 Simulation results of the CR for test data in different regions.

SELECTING OPTIMAL ML ALGORITHM

The optimisation and validation of learning models should be conducted separately for each of the three study regions, using two ML algorithms, and for all target indicators. This process would require 30 iterations (3 areas × 2 algorithms × 5 indicators).

Given that hyperparameter optimisation is computationally intensive, this process was performed exclusively on Region 1 data. This approach is justified on two grounds. First, all three regions share an identical feature space – the same five input variables (horizontal angle, vertical angle, inter-panel distance, roof elevation, and urban block), meaning the underlying data structure and complexity are consistent across regions. Second, a preliminary inspection of the simulation outputs confirmed that the statistical distributions of the target indicators across the three regions are comparable in range and variance, suggesting that the optimal model configuration identified for one region is transferable to the others. The hyperparameters identified for Region 1 were therefore applied directly to Regions 2 and 16, with validation metrics confirming their suitability (Tables 6 and 7).

Table 4 compares the two learning models based on the data from Region 1, across various indicators and using the optimal hyperparameters.

Table 4 Specifications of ML Model Hyperparameters.

ALGORITHM	TARGET INDICATOR	OPTIMAL HYPERPARAMETER
RF n_estimators and max_depth	HW	none, 45
	Electricity	none, 100
	Number of panels	none, 45
	CO_2 reduction	none, 100
ANN hidden-layer-sizes	HW	(100,200)
	Electricity	(250,250)
	Number of panels	(100,200)
	CO_2 reduction	(100,200,200)

VALIDATION OF LEARNING MODELS

At this stage, each model was trained and validated using the optimal hyperparameters for region 1. The accuracy evaluation metrics for each target indicator are presented in Table 5. Additionally, the time required for model training, as well as the target indicators for each model, have been compared.

Based on the comparison between the neural network and RF models using the R^2 recognition coefficient index and the time required for learning, it's evident that the RF model slightly outperforms the neural network model in terms of R^2. Additionally, the RF model demonstrates significantly better performance in terms of learning time compared to the neural network model. Therefore, based on these findings, the RF algorithm is applied in this research.

ALGORITHM	TARGET INDICATOR	LEARNING TIME (s)	VALIDATION					
			RMSE		MAE		R-SQUARE	
			TEST	VALIDATION	TEST	VALIDATION	TEST	VALIDATION
RF	HW	0.08	41,643.7	2,500.2	3,229.2	576.04	0.99	0.99
	Electricity	0.07	2,098.4	9.473	137.6	3.5	0.99	0.99
	Number of panels	0.05	0	0	0	0	1	1
	CO_2 reduction	0.1	8,100.81	627.38	731.09	231.98	0.99	0.99
ANN	HW	63.26	75,972.91	41,369.06	101,439.63	101,292.02	0.99	0.99
	Electricity	25.64	32,675.6	34,486.9	30,534.08	32,395.7	0.93	0.91
	Number of panels	37.6	51.6	26	20.03	17.4	0.99	0.99
	CO_2 reduction	50.28	14,301.5	7,901.5	6,709.2	5,916.8	0.99	0.99

Table 5 Performance of ML Models Based on Data from Region 1.

At this stage, each of the models has been trained and validated with optimal hyperparameters for region 2. Each of the accuracy evaluation indicators for each of the target indicators is displayed in Table 6. Also, the time required to learn the model has been measured in each target index. The duration of learning among the indicators varied from 0.05 to 0.09, and the recognition coefficient index among the target indicators was at least 0.99.

ALGORITHM	TARGET INDICATOR	LEARNING TIME (s)	VALIDATION					
			RMSE		MAE		R-SQUARE	
			TEST	VALIDATION	TEST	VALIDATION	TEST	VALIDATION
RF	HW	0.06	403.85	1,056.83	181.1	369.3	0.99	0.99
	Electricity	0.05	31.3	20.9	16.1	8.7	0.99	0.99
	Number of panels	0.08	0	0	0	0	1	1
	CO_2 reduction	0.09	218.92	263.41	131.79	131.41	0.99	0.99

Table 6 Performance of ML Models Based on Data from Region 2.

At this stage, each of the models has been trained and validated with optimal hyperparameters for region 16. Each of the accuracy evaluation indicators for each of the target indicators is displayed in Table 7. Also, the time required to learn the model has been measured in each target index. The duration of learning among the indicators was different from 0.05 to 0.95 and the recognition coefficient index was at least 0.99 among the target indicators.

ALGORITHM	TARGET INDICATOR	LEARNING TIME (s)	VALIDATION					
			RMSE		MAE		R-SQUARE	
			TEST	VALIDATION	TEST	VALIDATION	TEST	VALIDATION
RF	HW	0.09	752.53	1,663.24	363.66	597.99	0.99	0.99
	Electricity	0.07	38.27	23.31	17.94	4.5	0.99	0.99
	Number of panels	0.05	0	0	0	0	1	1
	CO_2 reduction	0.06	345.43	425.25	219.88	238.9	0.99	0.99

Table 7 Performance of ML Models Based on Data from Region 16.

MACHINE LEARNING RESULTS

The results of the simulation and the values predicted by the RF algorithm for each of the target indicators for each region have been compared and are presented in Figure 13.

SENSITIVITY ANALYSIS RESULTS

The correlation values between different data are visually represented in Figure 14, allowing for a clear understanding of the relationships between the variables. Each numerical value of the coefficients is clearly displayed, providing valuable insights into the strength and direction of the correlations observed in the data.

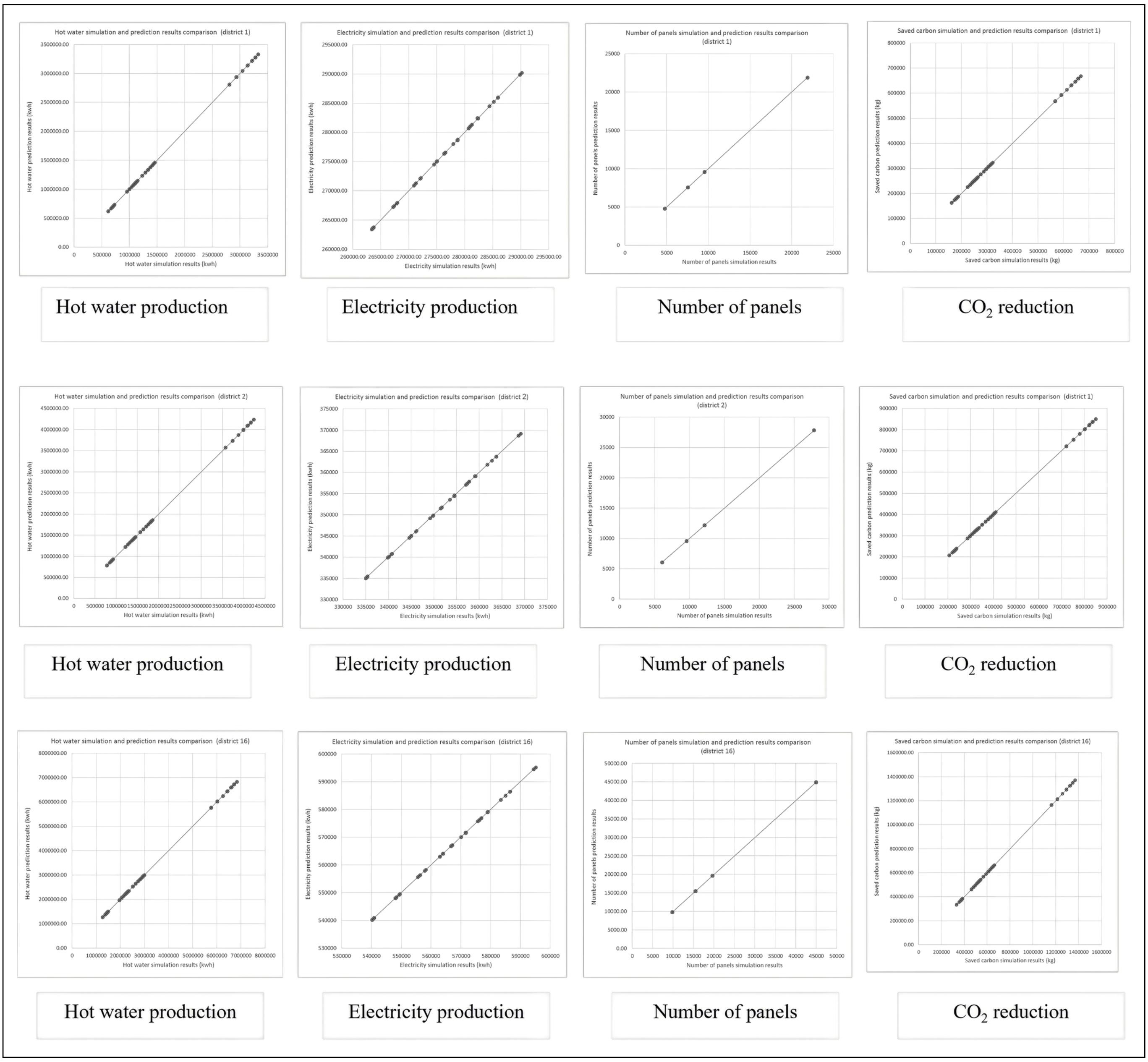

Figure 13 Comparison of Simulation and Predicted Results in Region 1, 2, 16.

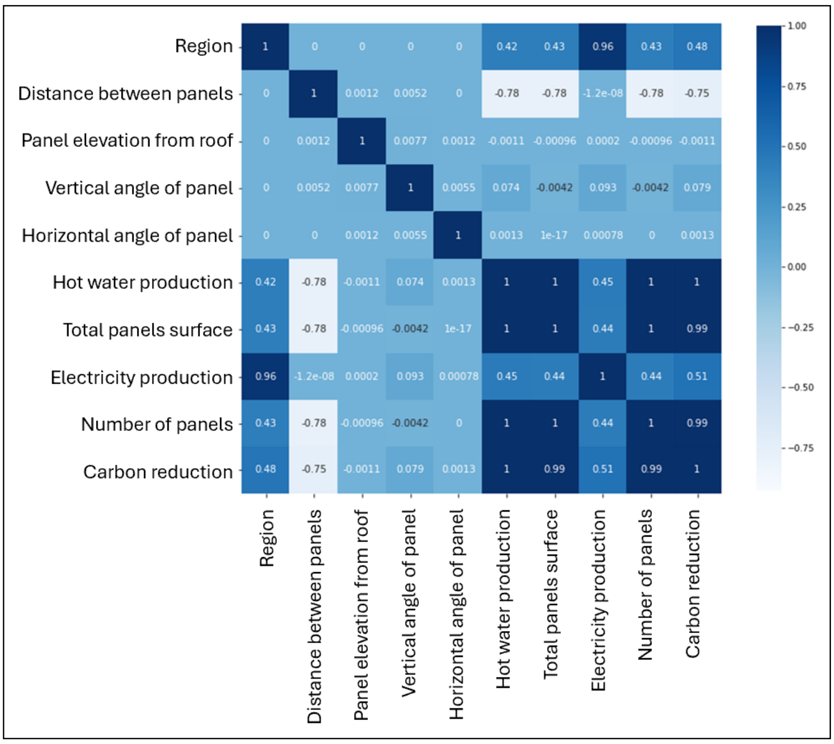

Figure 14 Overall sensitivity analysis.

DOI: 10.66506/essp.8-244-26

The correlation values shown in Figure 14, shed light on the factors influencing various aspects of the study:

- HWP: The most significant factor impacting HWP is the distance between panels, with a coefficient of -0.75. This indicates that decreasing the distance between panels (increasing panel density) leads to higher HWP. The density and height of buildings in the study region (coefficient: 0.42) also play a notable role. Additionally, the vertical angle of the panels (coefficient: 0.074) contributes to HWP, while the horizontal angle and distance of panels from the ceiling have lesser impacts.
- EP: The study region (coefficient: 0.96) is the most influential factor in EP, followed by the vertical angle of the panels (coefficient: 0.093). Other factors, such as the horizontal angle, distance of panels from the ceiling, and distance between panels, also contribute to EP, albeit to a lesser extent.
- Number of Panels: The distance between panels has the highest impact on the number of panels, with a coefficient of -0.78, indicating that increasing panel density reduces the number of panels required. The study region (coefficient: 0.43) is the next significant factor, followed by the vertical angle of the panels. The horizontal angle and distance of panels from the ceiling have minimal effects on the number of panels.
- Amount of CR: The distance between panels (coefficient: -0.75) has the greatest impact on CR, with higher panel density resulting in lower CR. The study region (coefficient: 0.48). Additionally, the vertical angle of the panels influences CR, while the horizontal angle and distance from the ceiling have minimal effects.

PROPOSED FRAMEWORK

The proposed framework (Figure 15) functions as a high-speed surrogate model designed to bypass the computational delays of traditional building performance simulations. It begins with the generation of a large-scale parametric dataset using GIS-based urban modelling and physical simulations, which captures the complex relationships between installation geometry and urban shading. This framework allows non-expert users to input specific urban configurations and receives near-instantaneous predictions for EP, HWP, and CR. Ultimately, it transforms a slow, specialist-led simulation process into a rapid, automated decision-support tool for optimising PVT deployments in dense city environments.

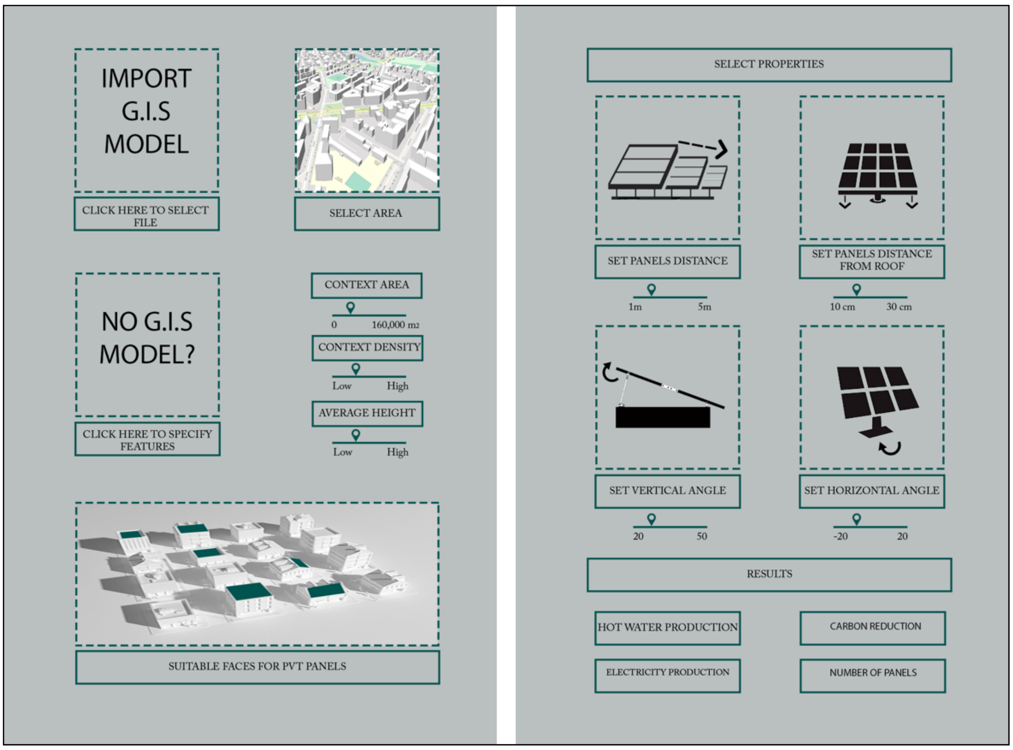

Figure 15 Proposed framework for evaluation of PVT panels in urban context.

DOI: 10.66506/essp.8-244-26

DISCUSSION

This study evaluates the development and performance of machine learning models for predicting the multi-faceted outputs of Photovoltaic-Thermal (PVT) systems within varied urban environments, drawing on the data and comparative literature provided in the sources.

INTEGRATED PERFORMANCE MODELLING

A significant contribution of this research is the integration of electrical, thermal, and environmental indicators into a single predictive framework. While much of the existing literature, such as Shin et al. (2022) and(Sulaiman et al. (2024) focuses primarily on building-level EP, this study expands the scope to include HWP and CR. By doing so, it addresses a gap identified in the literature where studies often prioritise single parameters. The results indicate that the RF algorithm is exceptionally well-suited for this multi-output task, achieving an R^2 accuracy of 0.91 to 1.0 across different indicators (Table 5).

ALGORITHMIC EFFICIENCY AND SELECTION

The comparison between ANN and RF highlights critical trade-offs in ML applications for renewable energy. While both models demonstrated high predictive accuracy ($R^2 \geq 0.91$), the RF algorithm was chosen as the superior model primarily due to its significantly shorter learning time. For instance, in Region 1 (high-rise buildings), the RF model's learning time for HWP was a mere 0.08 seconds, compared to 63.26 seconds for the ANN. This efficiency is vital for the proposed four-step software, which aims to facilitate decision-making for non-experts by providing rapid, data-driven results for urban PVT installations. Although this study focuses on Tehran, the proposed surrogate framework is transferable to other dense urban contexts, including European cities facing similar rooftop solar deployment challenges under the Energy Performance of Buildings Directive (EPBD) and the EU Renovation Wave policy frameworks (Hamdy et al. 2013; Pompei et al. 2026).

THE ROLE OF URBAN MORPHOLOGY

The findings underscore that urban density and building height are the most critical environmental determinants of PVT performance. The simulation results across Tehran's regions revealed that Region 16 (low-rise, high density) outperformed Region 1 (high-rise) by approximately 52% in EP and 50% in HWP. This disparity is directly attributed to shading effects caused by taller structures, confirming the findings of Wang et al. (2021) regarding the necessity of shading mitigation. Furthermore, sensitivity analysis showed that the study region itself had a 0.96 correlation coefficient with EP, indicating that the physical urban context is the single most influential factor for power generation.

OPTIMISATION OF INSTALLATION PARAMETERS

Beyond urban context, the study identifies specific installation variables that can be optimised to enhance system yield. Sensitivity analysis revealed that the distance between panels (density) is the most influential factor for HWP (coefficient: -0.75) and CR (coefficient: -0.75). This suggests that increasing panel density is a highly effective strategy for maximising thermal and environmental benefits. Additionally, while many studies in the literature review, such as Kassem and Othman (2022) rely on climatic variables like humidity and wind speed, this research demonstrates that geometric installation parameters (vertical/horizontal angles and roof elevation) are sufficient for high-accuracy performance modelling when integrated with urban GIS data.

CONCLUSION

In conclusion, the research demonstrates that machine learning, particularly Artificial Neural Networks (ANN) and Random Forest (RF), provides a robust methodology for predicting the complex outputs of Photovoltaic-Thermal (PVT) systems in urban environments. The study successfully integrated predictive models for EP, HWP, and CR, differentiating itself from prior research that often focuses on a single parameter.

Key findings from the analysis include:

- Algorithm Performance: While both models showed high accuracy, the RF algorithm was identified as the superior model due to its higher R^2 accuracy (0.91) and significantly shorter learning time compared to ANN.
- Urban Morphology Impact: The physical characteristics of urban districts, specifically building height and shading, are critical determinants of system performance. Panels in Region 16 (characterised by low-rise buildings and less shading) achieved approximately 52% higher EP and 50% higher HWP than those in the high-rise Region 1.
- Sensitivity Factors: Sensitivity analysis revealed that the distance between panels (density) is the most significant factor affecting HWP and CR, while the specific urban area has the greatest influence on EP.
- Practical Framework: The research culminated in a four-step software framework, incorporating GIS data and panel properties, designed to assist non-experts in optimising PVT installations for energy efficiency and urban sustainability.

While the current study provides a strong foundation using urban morphology in cold semi-arid climate, it acknowledges limitations regarding the range of input variables and geographic scope. Future research should aim to increase model generalisability by incorporating diverse weather conditions, different PVT module types, and the exploration of Deep Learning approaches to further enhance prediction accuracy.

DOI: 10.66506/essp.8-244-26

ABBREVIATIONS

AARD	absolute average relative deviation
AI	Artificial intelligence
ANFIS	adaptive neuro-fuzzy inference systems
ANN	Artificial Neural Network
BIPV	Building Integrated Photovoltaic
CFNN	Cascade Feed-forward Neural Network
CR	Carbon Reduction
DE	Differential Evolution
DNN	deep neural network
DT	decision tree
ELM	extreme learning machine
ENN	Elman neural network
EP	Electricity Production
ETR	extra tree regressor
GA	genetic algorithm
GBR	gradient boosting regressor
GE	Grammatical Evolution
GPR	Gaussian process regression
GR	Gaussian regression
HW	Hot water
HWP	Hot Water Production
KNN	K nearest neighbours
KNR	k-neighbours regressor
LASSO	least absolute shrinkage and selection operator
LGBM	Light gradient boosting machine
LR	linear regression
LS-SVR	Least Squares Support Vector Regression
LSTM	Long Short-Term Memory
MAPE	Mean absolute percentage error

MFFNN	Multilayer Feed-Forward Neural Network
ML	Machine learning
MLP	multilayer perceptron
MR	multivariate regression
MSE	Mean Squared Error
NMSE	Normalised Mean Squared Error
nRMSE	Normalised root mean square error
PR	polynomial regression
PV	Photovoltaic
PVT	Photovoltaic thermal
QSVM	quadratic support vector machine
RBFNN	Radial Basis Function Neural Network
RELU	Rectified Linear Unit
RF	random forest
RMSE	Root-mean-square deviation
SDA	similar day analysis
SVM	support vector machine
SVMR	support vector machine regression
SVR	support vector regression
XGBoost	extreme gradient boosting

DATA ACCESSIBILITY STATEMENT

The parametric dataset of 1,575 installation scenarios generated and analysed during this study, along with the trained Random Forest model and associated Python scripts, are available at: https://doi.org/10.5281/zenodo.19696118. The GIS shapefiles for the three Tehran urban districts were obtained from Tehran Municipality and are subject to their data sharing policies. The meteorological data used for simulation are based on the publicly available EnergyPlus Weather (EPW) file for Tehran (2020), accessible at https://energyplus.net/weather.

FUNDING INFORMATION

This research was part funded by the Sustainable Energy Authority of Ireland, RD&D PREDICT project [Grant No. 23/RDD/1046] and the MaREI, the Research Ireland Centre for Energy, Climate and Marine [Grant No. 12/RC/2303-P2].

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work, including conceptualisation, methodology, data collection and analysis, writing, and review of the manuscript.

ECEEE PAPER ID

8-244-26

AUTHOR AFFILIATIONS

Alireza Nazeri orcid.org/0009-0003-2201-0592
School of Mechanical Engineering, Technological University Dublin, Dublin, Ireland; Irish Building Stock Observatory, Dublin, Ireland

Ciara Ahern orcid.org/0000-0001-8417-5964
Senior Lecturer, School of Mechanical Engineering, Technological University Dublin, Dublin, Ireland; Irish Building Stock Observatory, Dublin, Ireland; MaREI, RI Centre for Energy, Climate and Marine, Cork, Ireland

REFERENCES

Alghamdi, H, Maduabuchi, C, Albaker, A, Almalaq, A, Alsuwian, T and **Alatawi, I.** 2023. 'Machine Learning Performance Prediction of a Solar Photovoltaic-Thermoelectric System with Various Crystalline Silicon Cell Types'. *International Journal of Energy Research* 2023. DOI: https://doi.org/10.1155/2023/1990593

Asiedu, ST, Nyarko, FKA, Boahen, S, Effah, FB and **Asaaga, BA.** 2024. 'Machine learning forecasting of solar PV production using single and hybrid models over different time horizons'. *Heliyon,* 10(7). DOI: https://doi.org/10.1016/j.heliyon.2024.e28898

Biyik, E. et al. 2017. 'A key review of building integrated photovoltaic (BIPV) systems'. *Engineering Science and Technology, an International Journal,* 20(3): 833–858. DOI: https://doi.org/10.1016/j.jestch.2017.01.009

Cao, Y, Kamrani, E, Mirzaei, S, Khandakar, A and **Vaferi, B.** 2022. 'Electrical efficiency of the photovoltaic/thermal collectors cooled by nanofluids: Machine learning simulation and optimization by evolutionary algorithm'. *Energy Reports,* 8: 24–36. DOI: https://doi.org/10.1016/j.egyr.2021.11.252

Chaibi, Y, Malvoni, M, El Rhafiki, T, Kousksou, T and **Zeraouli, Y.** 2021. 'Artificial neural-network based model to forecast the electrical and thermal efficiencies of PVT air collector systems'. *Cleaner Engineering and Technology,* 4. DOI: https://doi.org/10.1016/j.clet.2021.100132

Chiang, W, Permana, I, Wang, F, Chen, H and **Erdenebayar, M.** 2022. 'Experimental investigation for an innovative hybrid photovoltaic/Thermal (PV/T) solar system'. *Energy Reports,* 8: 910–918. DOI: https://doi.org/10.1016/j.egyr.2022.10.264

Dev, A, Kumar, R and **Saini, RP.** 2022. 'Experimental evaluation of performance of a hybrid solar photovoltaic (PV/T) panel integrated with effective cooling solutions with water base nanofluids and phase change materials'. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects,* 44(3): 7287–7302. DOI: https://doi.org/10.1080/15567036.2022.2107732

Elsaraiti, M and **Merabet, A.** 2022. 'Solar Power Forecasting Using Deep Learning Techniques'. *IEEE Access,* 10: 31692–31698. DOI: https://doi.org/10.1109/ACCESS.2022.3160484

Gharaee, H, Erfanimatin, M and **Bahman, AM.** 2024. 'Machine learning development to predict the electrical efficiency of photovoltaic-thermal (PVT) collector systems'. *Energy Conversion and Management,* 315. DOI: https://doi.org/10.1016/j.enconman.2024.118808

Gonçalves, JE, van Hooff, T and **Saelens, D.** 2020. 'Understanding the behaviour of naturally-ventilated BIPV modules: A sensitivity analysis'. *Renewable Energy,* 161: 133–148. DOI: https://doi.org/10.1016/j.renene.2020.06.086

Hamdy, M, Hasan, A and **Siren, K.** 2013. 'A multi-stage optimization method for cost-optimal and nearly-zero-energy building solutions in line with the EPBD-recast 2010'. *Energy and Buildings,* 56: 189–203. DOI: https://doi.org/10.1016/j.enbuild.2012.08.023

Jing, J, Zhou, Y, Wang, L, Liu, Y and **Wang, D.** 2023. 'The spatial distribution of China's solar energy resources and the optimum tilt angle and power generation potential of PV systems'. *Energy Conversion and Management,* 283: 116912. DOI: https://doi.org/10.1016/j.enconman.2023.116912

Kabilan, R. et al. 2021. 'Short-Term Power Prediction of Building Integrated Photovoltaic (BIPV) System Based on Machine Learning Algorithms'. *International Journal of Photoenergy,* 2021. DOI: https://doi.org/10.1155/2021/5582418

Kassem, Y and **Othman, AA.** 2022. 'Selection of most relevant input parameters for predicting photovoltaic output power using machine learning and quadratic models'. *Modeling Earth Systems and Environment,* 8(4): 4661–4686. DOI: https://doi.org/10.1007/s40808-022-01413-7

Mohana, M, Saidi, AS, Alelyani, S, Alshayeb, MJ, Basha, S and **Anqi, A.E.** 2021. 'Small-scale solar photovoltaic power prediction for residential load in saudi arabia using machine learning'. *Energies,* 14(20). DOI: https://doi.org/10.3390/en14206759

Patel, A. et al. 2022. 'A Practical Approach for Predicting Power in a Small-Scale Off-Grid Photovoltaic System using Machine Learning Algorithms'. *International Journal of Photoenergy,* 2022. DOI: https://doi.org/10.1155/2022/9194537

Pham, VHS and **Tran, HD.** 2023. 'Research on applying machine learning models to predict the electricity generation capacity of rooftop solar energy systems on buildings'. *Asian Journal of Civil Engineering,* 24(8): 3413–3423. DOI: https://doi.org/10.1007/s42107-023-00722-1

Pompei, L, Massulli, AR, Vespasiano, D and **Lo Basso, G.** 2026. 'Towards the EPBD and ETS2 Mandates: Renewable Energy-Driven Retrofit of a Northern Hotel in Italy'. *Energies,* 19(3). DOI: https://doi.org/10.3390/en19030707

Rojek, I, Mikołajewski, D, Moroziński, A and **Macko, M.** 2023. 'Machine Learning- and Artificial Intelligence-Derived Prediction for Home Smart Energy Systems with PV Installation and Battery Energy Storage'. *Energies,* 16(18). DOI: https://doi.org/10.3390/en16186613

Scott, C, Ahsan, M and **Albarbar, A.** 2023. 'Machine learning for forecasting a photovoltaic (PV) generation system'. *Energy,* 278. DOI: https://doi.org/10.1016/j.energy.2023.127807

Serrano-Luján, L, Toledo, C, Colmenar, JM, Abad, J and **Urbina, A.** 2022. 'Accurate thermal prediction model for building-integrated photovoltaics systems using guided artificial intelligence algorithms'. *Applied Energy,* 315. DOI: https://doi.org/10.1016/j.apenergy.2022.119015

Shin, WG, Shin, JY, Hwang, HM, Park, CH and **Ko, SW.** 2022. 'Power Generation Prediction of Building-Integrated Photovoltaic System with Colored Modules Using Machine Learning'. *Energies,* 15(7). DOI: https://doi.org/10.3390/en15072589

Suanpang, P and **Jamjuntr, P.** 2024. 'Machine Learning Models for Solar Power Generation Forecasting in Microgrid Application Implications for Smart Cities'. *Sustainability,* 16(14): 6087. DOI: https://doi.org/10.3390/su16146087

Sulaiman, MH, Jadin, MS, Mustaffa, Z, Daniyal, H and **Mohd Azlan, MN.** 2024. 'Short-term forecasting of rooftop retrofitted photovoltaic power generation using machine learning'. *Journal of Building Engineering,* 94. DOI: https://doi.org/10.1016/j.jobe.2024.109948

Tripathi, AK, Aruna, M, Elumalai, PV, Karthik, K, Khan, SA, Asif, M and **Rao, KS.** 2024. 'Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions'. *Case Studies in Thermal Engineering,* 59. DOI: https://doi.org/10.1016/j.csite.2024.104459

Wang, Z, Liu, J, Zhang, Y, Yuan, H, Zhang, R and **Srinivasan, RS.** 2021. 'Practical issues in implementing machine-learning models for building energy efficiency: Moving beyond obstacles'. *Renewable and Sustainable Energy Reviews,* 143: 110929. Available at: https://www.sciencedirect.com/science/article/pii/S1364032121002227.

Willmott, CJ and **Matsuura, K.** 2005. 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance'. *Climate Research*, 30(1): 79–82. Available at: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/.

Yousif, JH and **Kazem, HA.** 2021. 'Prediction and evaluation of photovoltaic-thermal energy systems production using artificial neural network and experimental dataset'. *Case Studies in Thermal Engineering,* 27. DOI: https://doi.org/10.1016/j.csite.2021.101297

Zazoum, B. 2022. 'Solar photovoltaic power prediction using different machine learning methods'. *Energy Reports,* 8: 19–25. DOI: https://doi.org/10.1016/j.egyr.2021.11.183

Zhang, S, Wang, J, Liu, H, Tong, J and **Sun, Z.** 2021. 'Prediction of energy photovoltaic power generation based on artificial intelligence algorithm'. *Neural Computing and Applications,* 33(3): 821–835. DOI: https://doi.org/10.1007/s00521-020-05249-z

Zhou, Y, Zhou, N, Gong, L and **Jiang, M.** 2020. 'Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine'. *Energy,* 204. DOI: https://doi.org/10.1016/j.energy.2020.117894

TO CITE THIS ARTICLE:
Nazeri, A and Ahern, C. 2026. A machine-learning surrogate model for optimising photovoltaic-thermal deployment in complex urban morphologies. *eceee Summer Study proceedings*, 1(1): 29, pp. 1–22. DOI: https://doi.org/10.66506/essp.8-244-26

Submitted: 20 March 2026
Accepted: 04 May 2026
Published: 31 July 2026

eceee Summer Study proceedings is a peer-reviewed open access journal published by European Council for an Energy Efficient Economy (eceee).