2023 European Conference on Computing in Construction
40th International CIB W78 Conference
Heraklion, Crete, Greece
July 10-12, 2023

ARTIFICIAL DATASET GENERATION TO ENHANCE THE DESIGN EXPLORATION OF RESIDENTIAL BUILDINGS THROUGH DATA-INFORMED ENERGY LOAD FORECASTING MODELS

Andrea Giuseppe di Stefano[1], Gabriele Masera[1], and Matteo Ruta[1]
[1]Politecnico di Milano, Milan, Italy

Abstract

This study aims to assist urban planners and building designers in taking informed decisions based on energy performance – simulating a real-world urban development scenario – using limited computational resources. In particular, this paper proposes a new approach that integrates existing studies on building loads forecasting by using a Generative Adversarial Network (GAN) generated dataset based on significant geometrical parameters. This overcomes the needs for large datasets – often difficult to access.

The results demonstrate that the data-driven approaches have addressed the buildings' load predictions with a reasonable accuracy while significantly reducing the calculation time required.

Introduction

To avoid dangerous anthropogenic interference with the climate system, two mitigation measures are possible: reducing greenhouse gas (GHG) emissions and enhancing greenhouse gas sinks. However, there are strong reasons for favouring the former over the latter[1]. The risks raised by a large-scale deployment of negative emissions technologies are much more significant than the issues raised by replacing fossil fuels with renewables. Although negative emissions projects have become necessary because of the low remaining global carbon budget, the first imperative today remains to reduce global emissions rapidly and drastically through a global energy transition (Bourban, 2022).

The building sector accounts for roughly 40% of the total energy consumption and 38% of the CO_2 emissions in the European Union (Saheb et al., 2015). On a global scale, the energy savings potential is estimated to be 53 Hexajoules annually by 2050 (United Nations Environment Programme, 2022), and building designers play a vital role in realising this huge energy savings potential.

Nowadays, architects and engineers use building performance simulations (BPS), abstracting real-world evidence, to support informed decisions to assess and reduce the environmental impact of buildings and meet strict requirements related to indoor climate and performance objectives. To find possible solutions, the design team must vary many design parameters such as building geometry, insulation thickness, glazing properties, and HVAC systems. However, the variation of these factors constitutes an enormous multi-dimensional design space, generating a multi-scale, interdisciplinary, complex problem to be solved. Regarding the Architecture, Engineering and Construction (AEC) sector, the IEA ANNEX-30 research shows that the choices about critical design parameters are determined in the early design stage, and more than 40% of the building energy-saving potential comes from the early design stage (Attia et al., 2013). Therefore, it is necessary to optimise the critical design decisions from the beginning of the project to improve building performance (Lin et al., 2021).

Data-informed building performance simulations

BPS is a powerful physics-based method for predicting a building's dynamic behaviour, renewable energy sources (RES) integration, and the building's sustainability intrinsic criteria harmonisation (Olu-Ajayi et al., 2022). Hence, synergetic implementation of the BPS, energy efficiency and RES integration is the only way to realise sustainable buildings and approach carbon-neutral city planning without omitting user behaviour. Accurate load forecasting is the premise of reasonable generation, transmission, and energy distribution arrangement at a city scale. Improving load forecasting accuracy is conducive to proper operation mode and maintenance plan in a power system or microgrid to reduce operational costs and improve the benefits of the power system or microgrid (Hou et al., 2022).

[1] According to Geden (2016), "By establishing the idea of negative emissions [into carbon budgets, during the IPCC's fifth assessment cycle], climate researchers have helped, unintentionally, to mask the lack of effective political mitigation action," because including carbon dioxide removal (CDR) in the carbon budget allows decision-makers to circumvent the original constraints on global emissions, while claiming that they are bringing climate change under political control (Geden, 2016).

Anyway, in urban scenarios, running thousands of simulations is an obstacle to the widespread adoption of design space exploration, uncertainty analysis, sensitivity analysis, and optimisation. Worse yet, thousands of simulations may be necessary to thoroughly explore the high-dimensional design space formed by the many design parameters. This computational issue may be overcome by creating fast metamodels (Østergård et al., 2018) using machine learning (ML) and artificial intelligence (AI) based tools. However, there is still a lack of methods, algorithms and tools to support building performance optimisation in the early design stage.

Essential for the development of a solid evidence base for the use of ML-based BPS (metamodels) is data empirically derived from large populations representing the real-world conditions of complex building stock. Still, for the most part, even basic information about energy demand in buildings, e.g., trends and patterns, along with simple descriptions of population and stock segmentations, is limited or simply lacking (Skea, 2012; Summerfield and Lowe, 2012).

Supporting the development of evidence-based data for the energy performance of buildings requires having access to different levels of information, from high-level aggregate ecological studies (i.e. using small area statistics), cross-sectional studies of individual units of observations (people, households, premises, meters, etc.), and exploratory studies. The risk is that without detailed data collection and storage, longitudinal analysis or systematic reviews of research findings is not viable to support project-by-project learning (Hamilton et al., 2015).

However, to cope with the lack of data and, at the same time, highlight the importance of data gathering, large-scale analyses can be conducted using artificial datasets. Artificial datasets consist of a certain amount of data derived from simulations (conducted using traditional methods) or ML approaches such as Generative Adversarial Networks (GAN), which generate data from a small dataset. In both cases, these data are structured in such a way as to have consistency between features (input data) of the different models analysed.

This paper analyses the application of ML-based BPS for predicting cooling loads based on an artificial dataset generated with a tabular GAN for data generation.

The aim of this proof-of-concept is to demonstrate (*Objective 1*) the effectiveness of ML-based tools in terms of accuracy – baseline, and (*Objective 2*) the effectiveness of these tools trained on an artificial dataset generated with GANs.

Cooling loads prediction

When it comes to energy-efficient building design, the computation of the heating load (HL) and the cooling load (CL) is required to determine the specifications of the heating and cooling equipment needed to maintain comfortable indoor air conditions.

These parameters are one of the most impactful for energy consumption and can be considered key performance indicators in a building design.

To estimate the required cooling and heating capacities, designers need information about the characteristics of the building and the conditioned space, the climate, and the intended functional use. Using statistical and machine learning concepts has the distinct advantage that distilled expertise from other disciplines is brought into the BPS domain. Using these techniques makes it extremely fast to obtain answers by varying building design parameters once a model has been adequately trained. Moreover, statistical analysis can enhance our understanding by offering quantitative expressions of the factors that affect the quantity (or quantities) of interest that the building designer or architect may wish to focus on (Tsanas and Xifara, 2012). Due to their intrinsic extensive data-based calculation, these tools can also be applied to a multi-scale domain, enhancing the possibility to study interrelations between multiple buildings and better understand city-scale energy consumption.

In this study, CL has been associated with some geometric building variables such as relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution (Pessenlehner and Mahdavi, 2003; Schiavon et al., 2010; Wan et al., 2011), combined with external factors such as climate (Wan et al., 2011). Starting from those data, a statistical analysis has been provided to gain insight into the underlying properties of input and output variables, using categorical regression and state-of-the-art nonlinear and non-parametric statistical machine learning tools to map the input variables to CL.

Methods

To evaluate the applicability of GAN-generated dataset for the cooling loads forecasting, we first selected a reference dataset. Next, we defined the design variables. We then conducted ML-based simulations to establish a baseline. We cropped the existing dataset – keeping the same variables and input/output relationships – and used to train a GAN in order to generate a second dataset, akin to the first. Finally, we conducted the same ML-based simulation to compare the two models and evaluate their performances.

The process is showed below in **Figure 01**, with the dashed boxes indicating the next steps to be performed in future work.

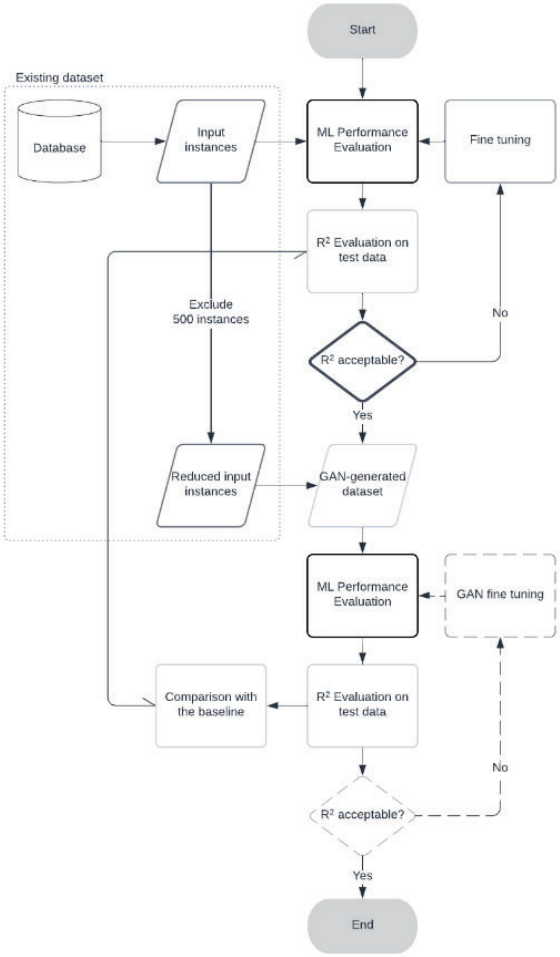

Figure 1: Methodology diagram

Case study

This section briefly summarises the data-driven statistical concepts and the ML techniques used to analyse the data. All the analyses are performed in a Jupyter Notebook using a Python environment. Some analytics libraries (such as Pandas, Numpy, Seaborn and Matplotlib) were used to process the data and obtain more readable results.

The used dataset is gathered from the Center for Machine Learning and Intelligent Systems data repository of the Bren School of Information and Computer Science (University of California), based on research by Tsanas and Xifara (Tsanas and Xifara, 2012). The data are based on a geometrical exploration starting from an elementary cube (3,5m × 3,5m × 3,5m) from which 12 building forms composed of 18 elements (elementary cubes) are generated.

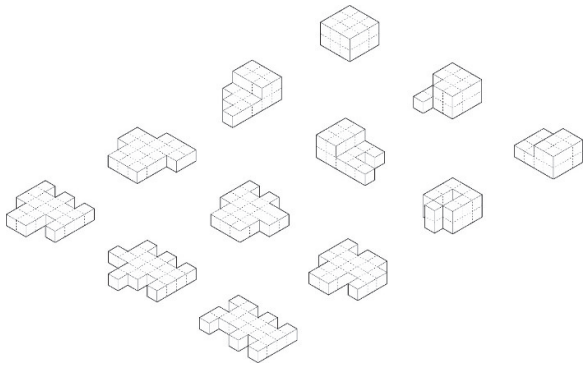

Figure 2: Representation of building forms generated from the combination of 18 elementary cubes

All the buildings have the same volume of 771 m^3 but different surface areas and dimensions. The materials used for each of the 18 elements are the same for all building forms. The selection was made by the most common materials in the building industry at the publication date, for a common building in Athens, Greece. Specifically, the associated U-values are: walls (1.78 W/m^2K), floors (0.86 W/m^2K), roofs (0.50 W/m^2K), windows (2.26 W/m^2K). The simulation assumes that the buildings are residential with seven persons and sedentary activity (70 W).

The internal design conditions were set as follows: clothing: 0.6 clo, humidity: 60%, air speed: 0.30 m/s, lighting level: 300 Lux. The internal gains were set to sensible (5 W/m^2) and latent (2 W/m^2), while the infiltration rate was set to 0,5 for air change rate with wind sensitivity 0.25 ach. For the thermal properties was used a mixed mode with 95% efficiency, with a thermostat range of 19–24 ˚C, 15–20 h of operation on weekdays and 10–20 h on weekends. Three types of glazing areas were used, expressed as percentages of the floor area: 10%, 25%, and 40%. Furthermore, five different distribution scenarios for each glazing area were simulated:

- uniform: 25% glazing on each side,
- north: 55% on the north side and 15% on each of the other sides,
- east: 55% on the east side and 15% on each of the other sides,
- south: 55% on the south side and 15% on each of the other sides,
- west: 55% on the west side and 15% on each of the other sides.

Finally, all shapes were rotated to face the four cardinal points. Thus, considering twelve building forms and three glazing area variations with five glazing area distributions each, for four orientations, 720 building samples. In addition, twelve building forms for the four orientations without glazing were considered. Therefore, in total, the dataset is based on 768 buildings samples. Each of the 768 simulated buildings can be characterised by the eight building parameters presented above.

As reported in the previous section, the data has 768 rows (instances) and 10 columns (dimension), of which 8 input values (features) and 2 output values. The input values are:

- Relative Compactness
- Surface Area - m²
- Wall Area - m²
- Roof Area - m²
- Overall height - m
- Orientation - 2:North, 3:East, 4:South, 5:West
- Glazing Area - 0%, 10%, 25%, 40% (of floor area)
- Glazing Area Distribution (Variance) - 1:Uniform, 2:North, 3:East, 4:South, 5:West

While the output:

- Cooling load – kWh

All the data can be summarised using simple diagrams, highlighting the differences between the 768 case-study buildings analysed.

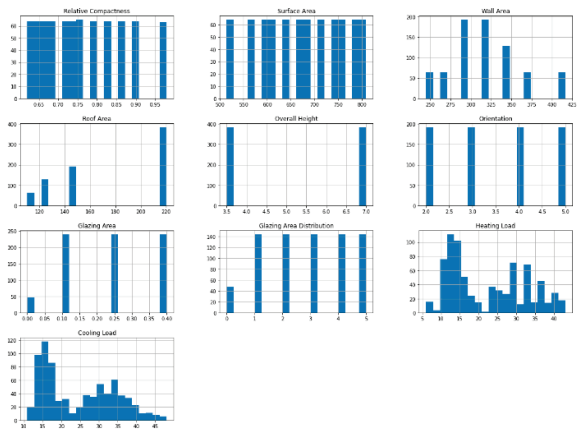

Figure 3: Value distribution of the input/output data

From a mathematical perspective, given N samples (here N=768) and M input variables (here M=8), we can construct a matrix $X \in R^{N \times M}$ which has the form of:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix} \quad (1)$$

This matrix is typically associated with a response variable vector $y \in R^{N \times 1}$ and we need to find the functional relationship f to relate X and y (here y is CL) such that $y = f(x)$. The tool that performs the functional mapping is commonly referred to as a learner in the machine learning literature.

Following this schema, the variables X (Relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution) and y (cooling load) have been defined. The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and supervised learning algorithms. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

For the ML-based simulation, the CatBoost regressor was used. CatBoost is an open-source machine learning algorithm (Prokhorenkova et al., 2017). It can work with diverse data types to help solve a wide range of problems. It yields state-of-the-art results without extensive data training typically required by other machine learning methods. CatBoost library is based on gradient boosting machine learning regression algorithm and is widely applied to multiple business challenges like fraud detection, recommendation items, and forecasting. It can return very good results with relatively few data, unlike other ML models that need to learn from massive amount of data. It also reduces the need for extensive hyperparameter tuning and lowers the chances of overfitting, leading to more generalised models.

Base data analysis

The first step of the process is the validation of the ML model based on the available dataset.

Gradient boosting is essentially a process of constructing an ensemble predictor by performing gradient descent in a functional space. It is backed by solid theoretical results that explain how strong predictors can be built by iteratively combining weaker models (base predictors) in a greedy manner. However, implementations of gradient boosting face the statistical issue of relying – after several steps of boosting - on the targets of all training examples. CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors. The CatBoost model use ordered boosting, avoiding target leakage, and a modified algorithm for processing categorical features, achieving better results over existing gradient boosted decision trees.

Using the CatBoost regression model, it has been possible to train the model in 4.79 seconds. Applying the cross-validation to the test subset, it is possible to note that the algorithm provides extremely accurate values in no time. The model was trained and validated on 33% of the data set, and the accuracy (R-squared value) for the prediction test was consistently above 90%. **Figure 7** shows the difference between actual and predicted data.

Table 1: Accuracy of the baseline model

Dataset	R-squared
Train dataset (y)	0.998
Test dataset (y)	0.991

GAN generated dataset

GAN is a deep learning generative technology. It contains two distinct ML models: generator and discriminator. The potential distribution of the raw data is explored through a confrontation strategy between the two models, thereby generating virtual samples consistent with the distribution of the raw base data (Mao et al., 2020). The generator is responsible for generating the synthetic data sample G(z) based on the original raw data and inputting them into the discriminator. The discriminator, on the other hand, is responsible has to distinguish the true and synthetic (generator-generated data) input samples (Jabbar et al., 2020). Generator (G) outputs the synthetic generated data samples, while the discriminator (D) outputs the sample discrimination rate, which, together, are converted into the objective optimisation function V(D, G) and then fed back to the generator and discriminator; iteratively, such process makes the generated data more and more realistic (Yu et al., 2022).

The GAN principle is depicted in **Figure 4** below.

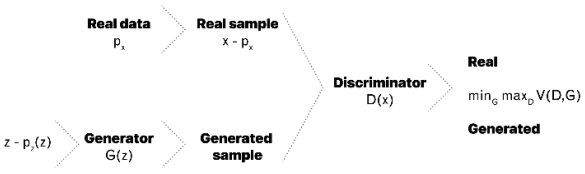

Figure 4: GAN functioning schema

The expression of objective optimisation process V(D, G) reads:

$$V(D,G) = E_{x\sim p_x}[logD(x)] + E_{z\sim p_z}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (2)$$

In **Equation 2**, x is the raw data, p_x is the distribution of x, z is the noise data, and p_z is the priori probability distribution of input noise variables.

The generator expects *V(D,G)* to be minimised, while the discriminator expects *V(D,G)* to be maximised, the process of which can be expressed as:

$$\min_G \max_D V(D,G) = E_{x\sim p_x}[logD(x)] + E_{z\sim p_z}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (3)$$

The generator and discriminator are fixed, respectively, to alternately and iteratively minimise and maximise V(D,G) until the final equilibrium is reached, thereby obtaining a GAN with better performance. At this time, the losses of the generator and discriminator are the same, and the artificial dataset is generated.

Artificial data generation and analysis

In this paper, we used the CTGAN library to generate the second artificial dataset. The preparation of the raw data consists of the original dataset, excluding 500 random raws, to give both a proper training set and a reasonably low number of simulations, like in a real-world scenario. The reduced raw dataset is then split into conditional and continuous columns to avoid physical errors, such as surfaces with an area <0.
According to this principle, the "orientation", the "glazing area", and the "glazing area distribution" parameters are considered discrete features, recurring the same steps as the original dataset. Lastly, an artificial dataset of 768 raws has been generated to be as coherent as possible with the original dataset. The figure below shows the accuracy of the GAN-generated dataset compared to the original one.

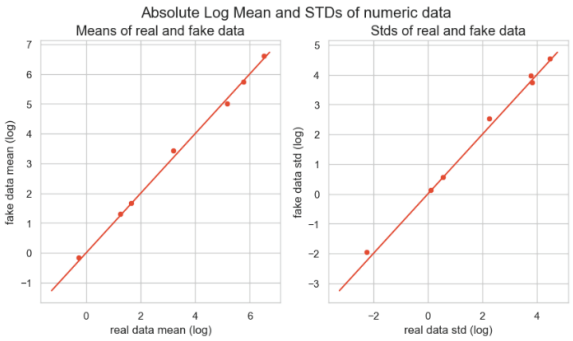

Figure 5: Distribution of means and standard deviation of real and synthetic data

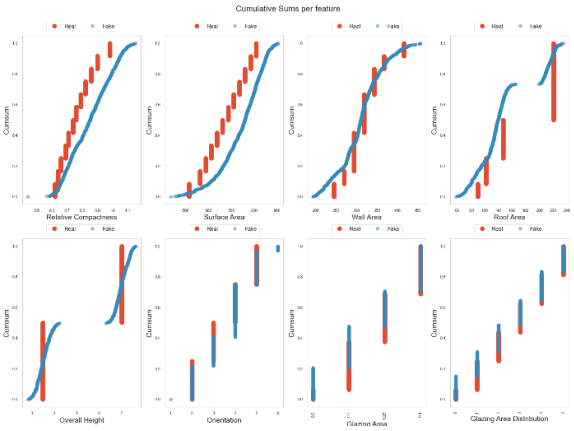

Figure 6: Real and synthetic data distribution for each feature

Lastly, the figure below shows the new distribution of the values from the GAN-generated dataset.

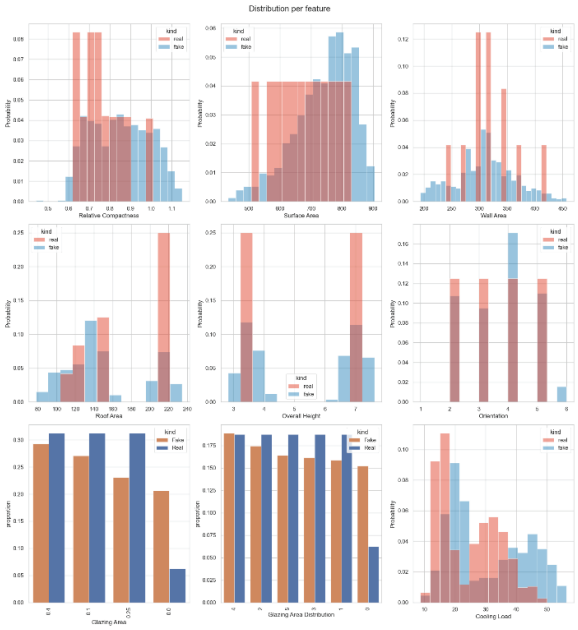

Figure 7: Value distribution of the input/output data, real and synthetic

Once the new dataset has been generated, in order to prove the accuracy of the model trained with the generated data against the baseline, the same ML model (CatBoost regressor) has been used.

After the neural network has been trained on the artificial dataset, the model is compared with the model trained on the original dataset. Finally, the outputs of the two models are compared so that the method's effectiveness can be assessed.

As expected, the effectiveness of the neural network based on the artificial dataset is significantly lower than the former, with an accuracy of around 40%. However, analysing the comparison graph, it is possible to see that the peaks (both positive and negative) are consistent and that the average consumption is in line with reality. This discrepancy may be dictated by the structure of the source data, which being variations of 12 buildings only, show repeating patterns, which is not absorbed in the artificial dataset. Furthermore, it is crucial to emphasise that both models were not normalised in order to reduce outliers or any values that could distort the overall behaviour of the model. This choice was made to test the feasibility of the approach in its crudest state. It is believed, therefore, that more accurate and usable results can be obtained after normalising the data and exploring the generation models in greater depth.

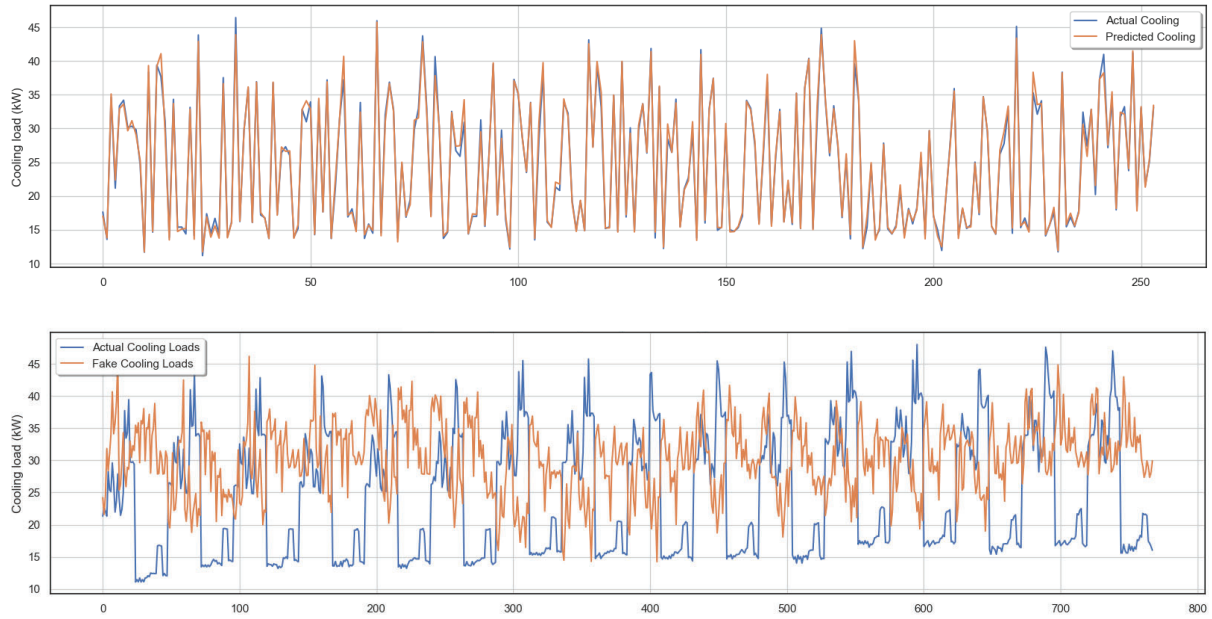

Figure 8-9: Comparison between predicted and actual cooling load (top), Comparison between the model trained with real and synthetic datasets (bottom)

Conclusions and outlook of future work

The test results demonstrate in practical terms how the ML-based tool responds (*Objective 1*) with a very high degree of accuracy (>90%) compared to the baseline calculated with traditional BPS methods. However, the results obtained with the GAN-generated dataset are not sufficient to guarantee the same accuracy level (*Objective 2*). The main issue can be found in the regression based on the artificial dataset, and not in the generation of the artificial dataset itself. This can be proven by looking at the comparison of values between raw and generated data that are consistent among all the features.

Future work will include further analyses (such as dataset normalisation, scaling, and different ML models comparison) on the application of ML algorithms to the artificial dataset, trying to overcome potential problems due to the double artificial modelling.

The outcomes of this paper confirm the potential of ML-based BPS for the exploration and optimisation of a significant design space in a limited timeframe. These results can be of great relevance in the hypothesis of early-stage design evaluations for the design of new buildings in an existing urban context, guaranteeing the possibility to evaluate different geometries in reduced timescales and maximise their performance.

These tools can be further explored and applied to the simultaneous analysis of multiple buildings (or variations of buildings) to rapidly assess and optimise the design of new urban or neighbourhood developments, taking into account the energy needs both of individual buildings and of the aggregate. In this scenario, ML-based BPS represents a fundamental step forward for net zero-carbon developments such as the ones defined in the European Commission's "100 EU Cities" for 100 climate-neutral and smart cities by 2030, decoupling the analysis needed to achieve decarbonisation targets from the large amount of time and computational effort required by traditional methods.

References

Attia, S., Hamdy, M., O'Brien, W., Carlucci, S., 2013. Assessing gaps and needs for integrating building performance optimization tools in net zero energy buildings design. Energy Build. 60, 110–124. https://doi.org/10.1016/j.enbuild.2013.01.016

Bourban, M., 2022. Ethics, Energy Transition, and Ecological Citizenship, in: Comprehensive Renewable Energy. Elsevier, pp. 204–220. https://doi.org/10.1016/B978-0-12-819727-1.00030-3

Geden, O., 2016. The Paris Agreement and the inherent inconsistency of climate policymaking. WIREs Clim. Change 7, 790–797. https://doi.org/10.1002/wcc.427

Hamilton, I., Oreszczyn, T., Summerfield, A., Steadman, P., Elam, S., Smith, A., 2015. Co-benefits of Energy and Buildings Data: The Case For supporting Data Access to Achieve a Sustainable Built Environment. Procedia Eng. 118, 958–968. https://doi.org/10.1016/j.proeng.2015.08.537

Hou, H., Liu, C., Wang, Q., Wu, X., Tang, J., Shi, Y., Xie, C., 2022. Review of load forecasting based on artificial intelligence methodologies, models, and challenges. Electr. Power Syst. Res. 210, 108067. https://doi.org/10.1016/j.epsr.2022.108067

Jabbar, A., Li, X., Omar, B., 2020. A Survey on Generative Adversarial Networks: Variants, Applications, and Training. https://doi.org/10.48550/ARXIV.2006.05132

Lin, B., Chen, H., Liu, Y., He, Q., Li, Z., 2021. A preference-based multi-objective building performance optimization method for early design stage. Build. Simul. 14, 477–494. https://doi.org/10.1007/s12273-020-0673-7

Mao, Y., He, Q., Zhao, X., 2020. Designing complex architectured materials with generative adversarial networks. Sci. Adv. 6, eaaz4169. https://doi.org/10.1126/sciadv.aaz4169

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., Ajayi, S., 2022. Machine learning for energy performance prediction at the design stage of buildings. Energy Sustain. Dev. 66, 12–25. https://doi.org/10.1016/j.esd.2021.11.002

Østergård, T., Jensen, R.L., Maagaard, S.E., 2018. A comparison of six metamodeling techniques applied to building performance simulations. Appl. Energy 211, 89–103. https://doi.org/10.1016/j.apenergy.2017.10.102

Pessenlehner, W., Mahdavi, A., 2003. Building morphology. transparence, and energy performance, in: Building Simulation. Presented at the Eighth International IBPSA Conference, Eindhoven, Netherlands, p. 8.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2017. CatBoost: unbiased boosting with categorical features. https://doi.org/10.48550/ARXIV.1706.09516

Saheb, Y., Bodis, K., Szabo, S., Ossenbrink, H., Panev, S., 2015. Energy renovation: the trump card for the new start of Europe. Publications Office, LU.

Schiavon, S., Lee, K.H., Bauman, F., Webster, T., 2010. Influence of raised floor on zone design cooling load in commercial buildings. Energy Build. 42, 1182–1191. https://doi.org/10.1016/j.enbuild.2010.02.009

Skea, J., 2012. Research and evidence needs for decarbonisation in the built environment: a UK case study. Build. Res. Inf. 40, 432–445. https://doi.org/10.1080/09613218.2012.670395

Summerfield, A.J., Lowe, R., 2012. Challenges and future directions for energy and buildings research. Build. Res. Inf. 40, 391–400. https://doi.org/10.1080/09613218.2012.693839

Tsanas, A., Xifara, A., 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. 8.

United Nations Environment Programme, 2022. Emissions Gap Report 2022: The Closing Window — Climate crisis calls for rapid transformation of societies. United Nations Environment Programme, Nairobi.

University of California, Irvine, n.d. Center for Machine Learning and Intelligent Systems | Bren School of Information and Computer Science. URL https://cml.ics.uci.edu/ (accessed 9.6.22).

Wan, K.K.W., Li, D.H.W., Liu, D., Lam, J.C., 2011. Future trends of building heating and cooling loads and energy consumption in different climates. Build. Environ. 46, 223–234. https://doi.org/10.1016/j.buildenv.2010.07.016

Yu, M., Li, L., Guo, Z., 2022. Model analysis of energy consumption data for green building using deep learning neural network. Int. J. Low-Carbon Technol. 17, 233–244. https://doi.org/10.1093/ijlct/ctab100