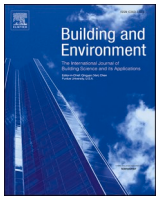
Building and Environment 250 (2024) 111126

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv



Examining the role of passive design indicators in energy burden reduction: Insights from a machine learning and deep learning approach

Siavash Ghorbany [a], Ming Hu [a,b,g,*], Siyuan Yao [c], Chaoli Wang [c], Quynh Camthi Nguyen [d], Xiaohe Yue [d], Mitra Alirezaei [e], Tolga Tasdizen [e], Matthew Sisk [f]

[a] Department of Civil and Environmental Engineering and Earth Sciences, College of Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA
[b] School of Architecture, Department of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, Notre Dame, IN, 46556, USA
[c] Department of Computer Science and Engineering, College of Engineering University of Notre Dame, Notre Dame, IN, 46556, USA
[d] Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, College Park, MD, USA
[e] Department of Electrical and Computer Engineering, Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA
[f] Lucy Family Institute for Data & Society, University of Notre Dame, Notre Dame, IN, 46556, USA
[g] Faculty of Architecture, Civil Engineering and Applied Arts, Academy of Silesia, Rolan 43, 40-555 Katowice, Poland

ARTICLE INFO

Keywords:
Average energy burden
Machine learning
Sustainable cities
Convolutional neural network
Computer vision

ABSTRACT

Passive design characteristics (PDC) play a pivotal role in reducing the energy burden on households without imposing additional financial constraints on project stakeholders. However, the scarcity of PDC data has posed a challenge in previous studies when assessing their energy-saving impact. To tackle this issue, this research introduces an innovative approach that combines deep learning-powered computer vision with machine learning techniques to examine the relationship between PDC and energy burden in residential buildings. In this study, we employ a convolutional neural network computer vision model to identify and measure key indicators, including window-to-wall ratio (WWR), external shading, and operable window types, using Google Street View images within the Chicago metropolitan area as our case study. Subsequently, we utilize the derived passive design features in conjunction with demographic characteristics to train and compare various machine learning methods. These methods encompass Decision Tree Regression, Random Forest Regression, and Support Vector Regression, culminating in the development of a comprehensive model for energy burden prediction. Our framework achieves a 74.2 % accuracy in forecasting the average energy burden. These results yield invaluable insights for policymakers and urban planners, paving the way toward the realization of smart and sustainable cities.

1. Introduction

The energy burden has been defined as the proportion of a household's gross income that is dedicated to energy expenses [1]. Out of all households in the United States, 38 %, equivalent to 46.5 million households, experience a significant energy burden. This ranges from those who pay over 6 % of their gross income on energy bills to those who face a severe energy burden, paying more than 10 % of their gross income [2]. Energy burden leads to the inability of a household to afford adequate energy sources to meet basic needs such as residential heating and cooling, cooking, cleaning, lighting, and using electrical appliances, which is a major issue in energy justice, especially for low-income families and disadvantaged communities who allocate a large proportion of their income to cover energy cost [2–4]. Low-income households bear energy burdens that are at least double that of the average household, while energy burdens for black households are 43 % higher than those for white (non-Hispanic) households [2]. Meanwhile, the high energy burden is associated with financial stress, which can lead to energy deprivation (e.g., cannot afford air conditioning or heating), consequently leading to health risks, especially climate change exacerbated risks [2,5]. For example, climate change is expected to significantly amplify both the frequency and intensity of extreme temperature events, leading to an increase in energy demand for cooling and heat-related health risks [6]. In general, climate change-induced cooling

* Corresponding author. Department of Civil and Environmental Engineering and Earth Sciences, College of Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA.
E-mail address: mhu1@nd.edu (M. Hu).

https://doi.org/10.1016/j.buildenv.2023.111126
Received 13 September 2023; Received in revised form 17 December 2023; Accepted 19 December 2023
Available online 30 December 2023

energy burden disproportionately affects marginalized communities and low-income households in ways: (1) low-income families often live in aging housing with deteriorated conditions, such as low insulated walls and roofs; (2) many households lack mechanical units/system for heating and/or air conditioning for cooling; (3) even if air conditioning units are present, many cannot afford the higher resulting utility bills.[3] There is an urgent and realistic need to find low-cost and low-technical threshold solutions to help low-income families prepare and adapt to climate change, introducing a higher energy burden and health risk.

Some studies have worked on the concept of low-carbon affordable technologies such as fans and in some cases it has shown these appliances are able to reduce the energy and improve the indoor comfort [7] and being beneficial to low income communities [8]. That said, researchers have identified passive design strategies, such as natural ventilation, as the preferred methods for constructing cost-effective and energy-efficient buildings [9–11], to adapt to the increasingly frequent and intensified extreme temperature conditions resulting from climate change involve minimizing energy costs and considering the possibility of power outages. Many passive strategies are inexpensive or even free, making them practical techniques for low-income communities. However, passive design strategies have been embraced as tried-and-true principles or traditional wisdom employed in historic construction practices. Still, there is a limited understanding and empirical evidence connecting passive design to energy burden, especially in developed countries. Research on passive design and its relation to energy demand and burden-is very limited at the neighborhood levels due to the paucity of high-quality data on passive design indicators. To our knowledge, there is no detailed passive design indicator dataset that can be used to study passive design's effect on mitigating energy burden at the neighborhood level (i.e., census tract level), which has hindered any thorough analysis on these variables' interactions and energy burden prediction. This urges the need for developing a platform that is able to capture these data and then analyze them with reliable precision. To fill the research gap, the goal of this project is: (1) develop a methodology and deep learning (DL) model to extract passive design indicators using Google Street View (GSV) images – a cutting-edge and readily available geographic data repository for conducting large-scale research; (2) examine whether passive design indicators are related to energy burden, with the Chicago metropolitan region being used as a testing bed, and (3) develop a comprehensive model to be able to explain the variances of average energy burden and predict it based on the passive design indicators and demographic data. The paper is organized as follows. Section 2 outlines the status of the art of passive design research and its gaps; section 3 explains the research methodology, data process, and sources; section 4 presents the findings, Section 5 outlines the research's significance, contribution, and limitations, and the conclusion is drawn in Section 6.

2. Current research and gaps

2.1. Defining passive design

This study specifically concentrates on passive design strategies, excluding active ones. Active systems rely on purchased or imported energy, such as electricity and gas, to artificially regulate indoor temperature through mechanical ventilation and air conditioning, while passive design harnesses natural ambient energy sources, such as sunlight and wind, to create a naturally comfortable indoor environment. These resources are considered as reliable renewable sources that are crucial to move towards sustainable cities. Passive design strategies, such as natural ventilation and harnessing solar heat, have been employed worldwide in traditional buildings to lower energy consumption and maintain comfortable indoor temperatures, predating the invention of modern mechanical heating and cooling systems [12,13]. Generally, passive design features do not consume energy during the operation and are low technology and low cost.

Previous studies found low-cost passive design strategies - such as utilizing natural ventilation and daylight in the appropriate form (e.g., south-orientating windows in northern hemisphere locations), are highly effective ways to reduce energy demand and, consequently, energy burden [10,14]. Our recent comprehensive review paper found that using combined passive strategies (e.g., natural ventilation) can achieve, on average, a cooling load reduction of 31 % and energy savings of 29 % in hot climate globally [15]. Another study also found when combining highly insulated building exterior walls and roof, the energy use of a house can be reduced used by 90 %, consequently reducing the energy burden [16]. The previous studies are found in developing countries, such as China [17–19], Nigeria [20], and Mexico [21], while very few studies are found in developed countries [15]. Moreover, the majority study was conducted on individual buildings as a case study, and very little research was found on the neighborhood level and linking passive design to energy burden.

2.2. Machine learning model and its application in energy burden prediction

Machine learning is a subfield of artificial intelligence in computer science that focus on the development of algorithms and statistical models that enable the computer system to learn from data and improve their performance on specific tasks [22,23]. Machine learning have been applied successfully across diverse research domains, from engineering, computer vision, finance, entertainment and computational biology [23–27]. Its remarkable efficiency in elucidating complex interrelationships between variables and making predictions has garnered significant attention [27–29]. Within the realm of statistical machine learning (ML), the methods can be categorized into supervised, unsupervised, and estimation (regression) methods [30,31]. Furthermore, it encompasses other techniques like semi-supervised learning, reinforcement learning, self-supervised learning, ensemble learning, and more, extending its scope [22,30,31]. These methods can be applied for various purposes, encompassing classification, clustering, and regression models [27,32].

Our research focuses on predicting the average energy burden, a crucial variable with significant implications for urban planning, especially in metropolitan regions affected by the urban heat island effect [33]. Given the nature of the average energy burden as a continuous numerical variable, our approach utilizes supervised estimation ML methods, known as regressions, to accurately capture and predict this variable [34].

Among these regression methods, decision tree regression [35], random forest regression [36], and support vector regression (SVR) [37] have gained prominence in prior research for predicting continuous numerical targets in diverse domains, including price prediction, energy demand estimation, and energy consumption forecasting [38–40]. SVR, in particular, has found application in energy-related areas, demonstrating comparable performance to Artificial Neural Network (ANN) models in estimating daily energy load requirements [39]. Moreover, ML-based regression models offer the flexibility to assess data within both linear and non-linear settings, employing various kernel functions, such as polynomial, sigmoid, and radial basis function (RBF) [41].

The energy domain has extensively utilized ML-based techniques, from predicting energy costs to ensuring food-energy security [42,43]. For instance, researchers leveraged Python programming to implement kernel guideline component analysis (KPCA) and Support Vector Machine (SVM) algorithms, achieving satisfactory results in predicting energy prices, considering a myriad of factors, including environmental, political, and supply-demand variables [42]. Other endeavors successfully predicted electricity load, achieving superior accuracy by combining SVR with the differential evolution (DE) algorithm [44]. Notably, the linear SVM method significantly enhanced sensor performance, leading to a remarkable 99 % accuracy rate in controlling light and optimizing energy usage [45]. Predicting electricity consumption,

incorporating house appliance data, weather conditions, and utilizing the Random Tree algorithm, has been instrumental in creating smart electricity consumption monitoring platforms, contributing to the vision of smart cities [40]. Many prior studies have concentrated on energy demand prediction in early-stage designs, employing ML methods such as ANN, SVM, and Random Forest, utilizing inputs like solar radiation values and building decomposition models [46]. For instance, Rao et al. conducted research on energy load forecasting for cluster microgrids in urban energy systems, identifying artificial neural networks, particularly with Levenberg–Marquardt optimization, as the most effective approach for accurate short-term load prediction [39]. However, these endeavors primarily revolved around daily load demand predictions based on solar and temperature data, with no specific focus on predicting the energy burden. In addition to load prediction, ML techniques have been instrumental in developing maintenance plans aimed at preventing power outages. Tree-based models, including the Random Forest, have yielded prediction accuracy exceeding 90 % in this context [47].

Despite these significant contributions in the field of energy prediction, a comprehensive model that encompasses passive design factors and demographic information to perceive and predict the average energy burden is notably absent. This gap in research is particularly relevant in understanding the amplified energy impact due to climate change in urban areas, with far-reaching implications for urban policy and residents' well-being [48,49].

3. Methodology

Fig. 1 illustrated this research includes four major steps including: (1) DL enabled computer vision model development, (2) energy burden data extraction and preparation, (3) database cleaning and merging, and (4) data analysis. In the first step, a DL-enabled computer vision model is developed based on a convolutional neural network (CNN) model that extracts the building's passive design indicators from GSV images. As

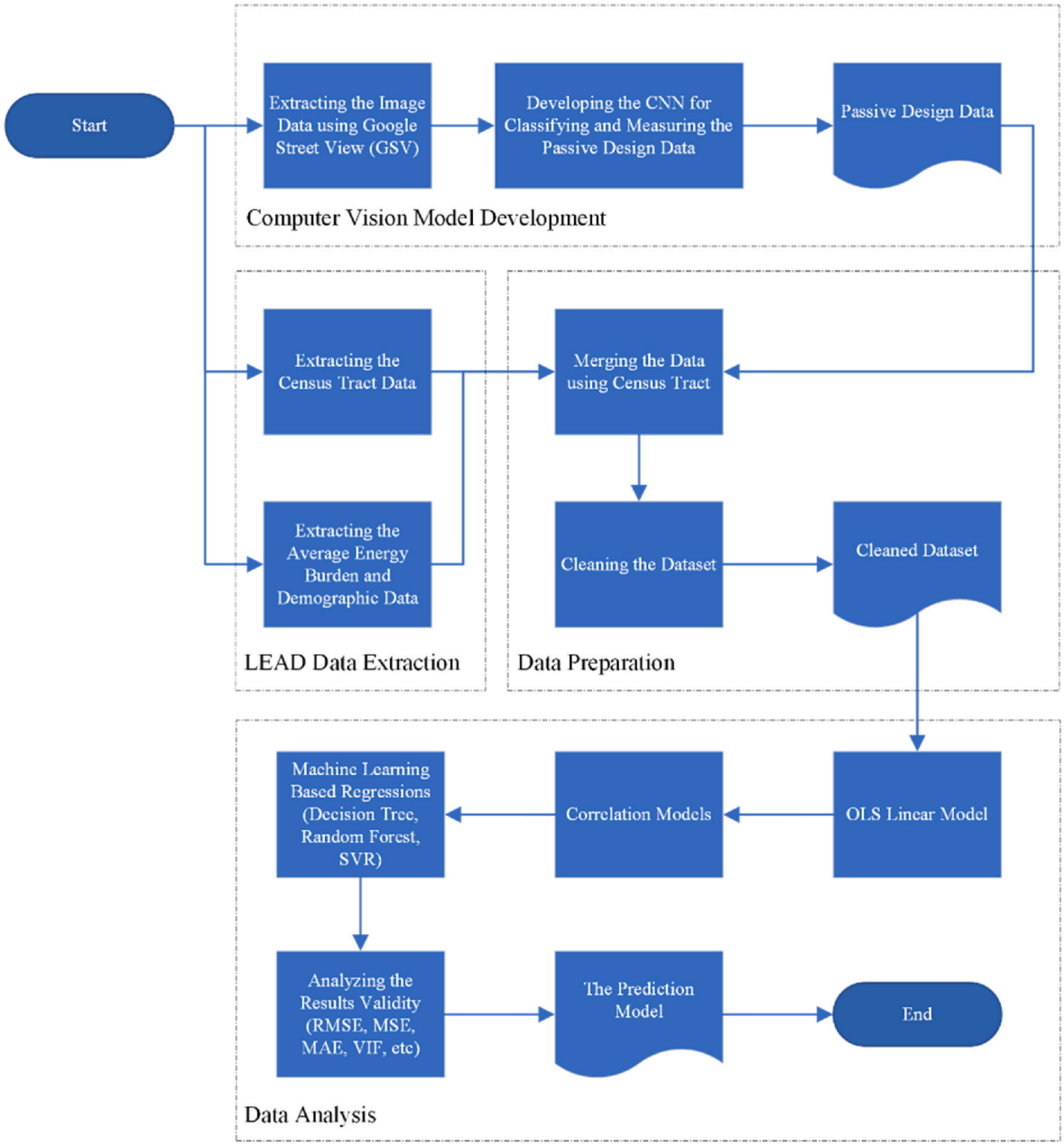


Fig. 1. The research methodology.

the second step, Chicago energy burden data are extracted from the 2018 Low-Income Energy Affordability Database (LEAD) at census tract level [50]. The third step includes aggregating the GSV and LEAD data based on the Census Tracts and cleaning the dataset for further analysis. The last step of the research starts by developing a multiple linear regression and correlation model as one of the basic data analysis models to first assess the association of passive design indicators with the average energy burden and, second, examine the suitability of this method as a predictive model. Subsequently, this study delved into an exploration of three of the most commonly employed ML regressions. It rigorously examined their performance in terms of error and accuracy when applied to demographic data and passive design characteristics (PDC) linked with the average energy burden. The objective was to identify the highly effective predictive model for energy burden forecasting. The validation and assessment of model fit involved the utilization of multiple measures, including the root mean square error (RMSE). These detailed steps are explained in the following sections.

3.1. Street view image collection

We acquired GSV image data through the GSV Image API. Our sampling strategy involved selecting locations at 100-m intervals. We collected GSV images from four distinct orientations (facing west, east, north, and south) for each set of coordinates (latitude and longitude). This approach was chosen to depict the housing quality comprehensively. About 227,000 images with 640 × 640 pixel resolution were downloaded for the Chicago metropolitan area, and the photos were from November 2019. Afterward, each of these images coded with coordinates was assigned to the related census tract Federal Information Processing System (FIPS) codes using geographic information system (GIS). After the passive design variables were extracted from each image, the values were aggregated to census-tract level for the next analysis step.

3.2. Housing passive design variables

Table 1 lists these passive design variables selected for this study as the optimal choices for illustrating the connection between housing conditions and energy demand. Fixed windows refer to the windows that cannot be opened, and non-fixed windows refer to that can be opened (operable windows). The selection of those variables is driven by the scientific findings from our published literature review, other published studies, and existing energy simulation tool input parameters. First, our published review paper on the effectiveness of passive design shows that natural ventilation is one of the most effective passive strategies for cooling energy load reduction [15]. To enable natural ventilation, non-fixed windows are a prerequisite condition. In addition to our review work, previous researchers also found the non-fixed window-to--wall ratio (WWR-Nonfixed) indicates access to natural ventilation, an energy use reduction mechanism for summer [51–53]. The fixed window-to-wall ratio indicates the potential thermal loss (due to its lower thermal property than insulated wall materials), and consequently, energy use increases for winter [19,54]. The external shading indicates direct solar radiation avoidance, which can be beneficial for summer cooling reduction but disadvantageous to solar heat gain in winter [53]. Thirdly, window types and window area and external shading are important input parameters in building energy models to predict the potential energy use [55]. Another reason for us to choose those passive design indicators is that we focused on the visual passive design features that can be extracted from GSV that can be scaled up and then applied to other geographic areas.

Table 1
Neighborhood sustainability indicators.

Passive Design	Indicators	Measurement Unit
Solar Optimization	Fixed window-to-wall ratio (WWR-Fixed)	Percentage
Natural Ventilation	Non-fixed window-to-wall ratio (WWR-Nonfixed)	Percentage
	External shading	Yes/No

3.3. DL-enabled Convolutional Neural Networks model

In this study, Convolutional Neural Networks (CNNs) were employed to process the GSV images and extract the passive design variables. CNNs belong to a category of multi-layer deep neural networks that have demonstrated their efficacy in the identification and extraction of features from data. These networks are widely applied for tasks such as image classification, object recognition, and the analysis of structured data arrays [56]. Convolutional Neural Networks (CNNs) are deep learning models capable of taking an input image, assigning weights to various features within the image, acquiring an understanding of these features, and distinguishing between them [57]. Some of the key applications of CNN application in building research include the estimation of building conditions (e.g., crack detection) [58–60], identifying abandoned property [61], and rooftop solar panel [62]. CNN offers an efficient, scalable, and dependable approach to transforming inputs into outputs by utilizing multiple layers that learn to detect various features within an image. To enhance model performance and mitigate overfitting issues, this project aimed to overcome the limitations associated with a relatively small dataset of labeled images. To achieve this, the models were pretrained on COCO dataset. The COCO training, validation, and test datasets collectively consist of over 200,000 images spanning 80 different object categories [63]. To optimize the model architecture, an COCO pretrained Faster-RCNN model [64] can be fine-tuned, involving adjusting the model's parameters, using a smaller amount of training data specific to the desired task. This approach enables substantial enhancements in model performance without demanding the extensive volume of training data and computational resources that were essential for training the original network [65].

In order to create an ML-enabled CNN model, we first created a training dataset through manual labeling/annotation. We used the COCO annotator to manually label 1290 images for four variables: Buildings, Fixed Windows, Non-fixed Windows, and External shading. As illustrated in Fig. 2, there were multiple boundaries labeled to represent multiple buildings in the same image such as blue and red boxes. "Buildings" identify the boundary of buildings in the images demonstrated with blue boxes. Fixed and Non-fixed Windows identify the boundary of different windows in the red box. Fixed and non-fixed windows indicate the natural ventilation potential [52,53]. After the building boundary and window boundary were identified, the ratio between Window to Wall was also calculated, as indicated by the numeric number in the image (e.g., 0.15). Lastly, External shadings were identified by the green box and assigned as binary labels (0,1) to indicate their presence (no or yes). In one image, there might be multiple shades (multiple buildings and multiple shades), and the final value used for the census tracts is the average value of these 0 and 1, equal to percentage units. Inter-rater agreement was close to 80 % for all indicators. Each image has multiple bounding boxes to detect the four variables (refer to Fig. 2). Next, we proceeded to randomly partition the image datasets into two distinct sets: a training set and a test set. The training set encompassed 75 % of the total labeled images, while the remaining 25 % were reserved for assessing the model's accuracy. We utilized these labeled sets to train an object detection model.

Fig. 3 showcases the architecture of our object detection system, which leverages a pre-trained Faster R–CNN model initially trained on the COCO dataset. This model was then fine-tuned using Google Street View (GSV) images.

The Faster R–CNN network is a powerful object detection architecture that consists of two essential components: the region proposal

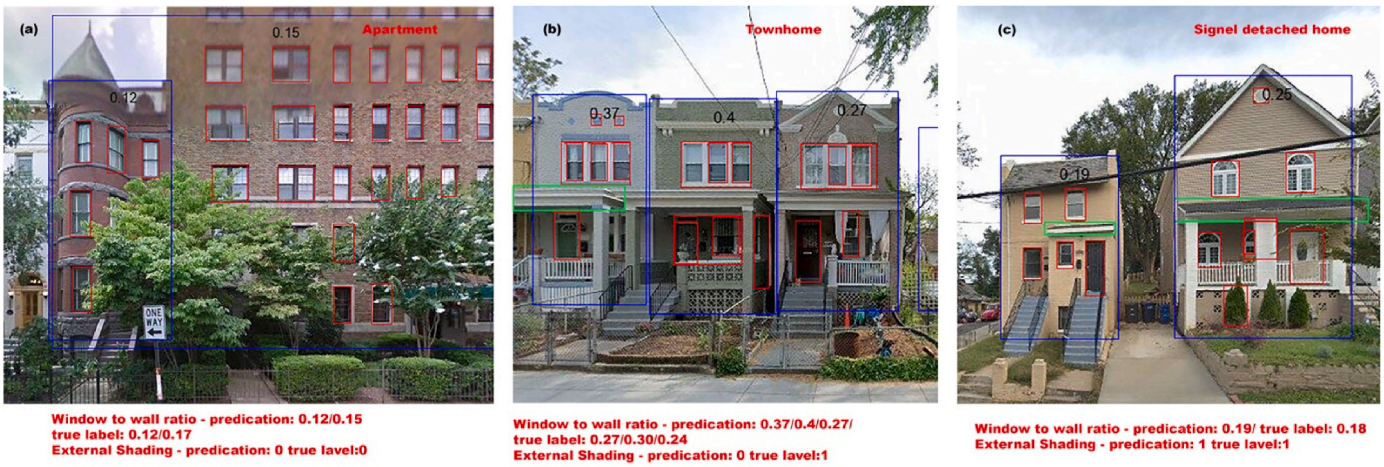


Fig. 2. Examples of processed GSV images for Fixed Windows, Non-fixed Windows, Buildings, and "External shading" indicators.

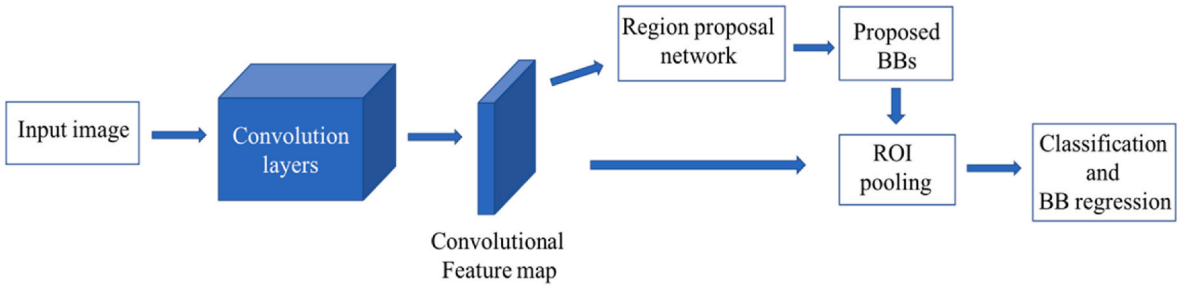


Fig. 3. Faster-RCNN model [64] that used for object detection. Each sample consists of an individual image along with associated labels and bounding boxes (BBs) that correspond to each indicator. The computer vision model consists of two parts: the region proposal network for suggesting regions with potential objects, and the classification and regression network that predicts both the object's class and its precise location.

network (RPN) and the subsequent object detection and classification part. The RPN network is responsible for proposing potential regions of interest within an image, suggesting where objects might be located. The second part of the network takes these proposals and performs bounding box detection and classification on each detected region. In our implementation, we leveraged a pre-trained Faster R–CNN model, initially trained on the COCO dataset. The weights of the network were initialized from the model trained on COCO dataset. We customized this model by replacing the classification head with one tailored for our specific task for predicting windows and buildings. The training process was done using Adam optimizer with a learning rate of $1e^{-5}$. It's important to note that the testing set remained concealed until we identified the best models through the training set. Subsequently, we utilized the testing set to assess the ultimate performance and quality of the model.

To obtain the indicators from Google Street View (GSV) images, we employed a pre-trained Faster R–CNN model on the COCO dataset and further fine-tuned it using our own labeled dataset. This dataset consists of GSV images with annotated bounding boxes specifically for detecting windows and buildings. During the fine-tuning process, the model was trained on the labeled training set to enhance its detection capabilities. For training, we utilized the Stochastic Gradient Descent (SGD) optimizer with a learning rate set to 0.003. This optimization approach was chosen to iteratively update the model's parameters and improve its performance on our task.

The model development dataset was divided into training and validation sets. The test set was deliberately kept unseen to the network during the training process to ensure an unbiased evaluation of the model's performance on new and unseen data. Also, to obtain a robust model and prevent overfitting, we implemented cross validation method. This process involved repeating the training and validation cycles by changing the subsets designated for training and validation.

In terms of evaluation and accuracy, we employed the mean Average Precision (mAP) as a key metric. The overall mAP for detecting windows and buildings is 0.74. Mean Average Precision is a common metric in object detection, calculated based on the intersection over union of the predicted bounding boxes and the ground truth bounding boxes. This metric provides a comprehensive assessment of the model's accuracy in localizing and classifying objects in the given images. We regarded the model saved in the concluding epoch as our ultimate model. The recognition task accuracy, which reflects the agreement between manually labeled images and computer vision predictions, was as follows: buildings (90.25 %), WWR-Fixed (81.25 %), WWR-Nonfixed (83.87 %), and External Shading (83.00 %).

3.4. Energy burden data and demographic and socioeconomic data

Census-level energy burden is quantified as the annual average housing energy expense divided by the annual average gross income of household [66], and represented in the Equation (1) [50]:

$$Energy\ Burden\ (E_b) = \frac{S}{G} \tag{1}$$

where S is the energy expenditures, G is the household income. Energy burden data was downloaded from the 2018 Low-Income Energy Affordability Database (LEAD), which is managed by the Department of Energy [50]. LEAD provides three primary metrics: energy burden, annual average housing energy costs, and housing counts. Energy burden is calculated as the average annual housing energy costs divided by the average annual household income. Monthly housing energy costs encompass household expenditures for electricity, gas (utility and

bottled), and other fuels, including fuel oil, wood, etc. Housing counts represent the number of occupied housing units. LEAD categorizes housing types into the following categories: single-family detached, single-family attached, 2-unit, 3–4 unit, and more than 4-unit apartments. These data are comprehensive, covering all 50 states, along with Washington D.C and Puerto Rico. They are accessible at multiple geographic levels, including national, county, city, and census tract levels [50].

Our analyses accounted for census tract, median housing age, poverty rate, percent Asian, percent Black, and percent Hispanic, and percentage of the population who are 65 or over. Covariate information was obtained from the American Community Survey 2018 5-year estimates.

3.5. Analytic approach

The GSV database was first merged with LEAD data at the census tract level to examine the dependent variable, Average Energy Burden relation with the independent variables, namely WWR-Fixed, WWR-Nonfixed, and External shading. Statistical analyses were implemented using Python programming focusing mainly on Pandas, NumPy, Matplotlib, SciPy, and Seaborn libraries. Python has been used and recommended as a robust and reliable tool for statistical analyses [67], large dataset handling, and machine learning analyses [27,68,69] in previous studies and also has been suggested for regression analyses [67], specifically ordinary Least Squares (OLS), which is the case of this research [69]. In order to provide an initiative insight into the dataset and its variables, Pearson and Kendall correlation were implemented to examine the variables influence and relationship to each other. Subsequently, we applied the Ordinary Least Squares (OLS) model to the dataset to calculate the relationships between passive design indicators derived from GSV data and energy burden. This analysis accounted for potential confounding factors, including racial/ethnic composition, housing age, and economic disadvantage. The regression model incorporated several covariates, encompassing the proportion of the overall population aged 65 to 74, the proportion of the total female population, the counts of Asian and Black or African American residents, the population living below the poverty line, the median construction year of buildings, the Hispanic population, and the population of individuals aged over 65. The selection of OLS was motivated by its capability to handle continuous data and effectively handle scenarios involving multiple independent variables [67,69], which aligns with the requirements of this research question. The OLS equation has been shown in Equation (2), where β' _denotes the ordinary least squares estimator, *X indicates* the matrix regressor variable X, *T demonstrates* the matrix transpose, and *y* stands for the vector of the value of the response variable.

$$= (XT\mathrm{X})^{-1}X^{T}\mathrm{y} \tag{2}$$

Moreover, the variance inflation factor (VIF) was calculated to prevent the multicollinearity and its consequent errors in the developed model [69]. VIF is a statistical metric used to evaluate multicollinearity in a regression study. When two or more independent variables in a regression model have a strong correlation with one another, this is known as multicollinearity, which can cause instability in the coefficient estimates and make it challenging to understand the relationships between the variables [69]. The VIF formula has been demonstrated in Equation (3), where R^2 denotes the coefficient of determination derived from regressing the focal independent variable against all other independent variables within the model.

$$VIF = \frac{1}{1 - R^2} \tag{3}$$

Following the implementation of the OLS method, other ML-based regression methods were also executed to achieve the most accurate model to describe and predict the average energy burden. This section showed how accurately the current variables can be used to predict the average energy burden. To achieve this, three supervised estimation ML methods, namely Random Forest Regression (RFR), Decision Tree Regression (DTR), and Support Vector Regression (SVR) were used. Then, grid search, as a well-renowned fine-tuning method in ML methods, was implemented to find the best result in each of these regression models [38,70]. To investigate the suitability and validity of the models, root mean square error (RMSE), mean average error (MAE), and mean squared error (MSE) were used as the recommended indicators to assess the machine learning models, regression models, and cause and result based models' errors and R-squared was used to indicate the model's accuracy [27,41,71,72]. The RMSE, MAE, and MSE are calculated based on Equations (4)–(6), respectively, where y_i^p is the estimated data, y_i is the predicted value, n is the sample size and $\bar{y}^p$ shows the mean value [68,73].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n}} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i^p - y_i| \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i^p - y_i)^2 \tag{6}$$

The choropleth maps are then created using ArcGIS Pro software (ESRI, Inc.) with geographical data sourced from US Census Bureau 2020 TIGER/Line geodatabases. The average energy burden is indicated as a percentage and visualized in three buckets as shown in Fig. 4: normal (1%–2%), energy poor (3%–5%), and energy poverty (6%–10 %). Similarly, we visualize WWR-Fixed, WWR-Nonfixed, and External shading in five buckets (See Figs. 4 and 5). The higher values are presented in deeper colors from the color scheme.

4. Results and findings

4.1. Descriptive statistics

Table 2 displays descriptive statistics of the passive design variables, demographics, and energy burden data summarized at the census tract level in the Chicago metropolitan areas. A total of 1441 census tract areas are included in the analysis. The average WWR for fixed windows is around 4.8 %, and for non-fixed operable windows is an average 0.5 %, with a maximum of approximately 15 %. The WWR is much lower than the current maximal ratio allowed by ASHRAE Standard 90.1, which is 40 % [74]. This study uses the median year built for each census tract, with a minimum value of the year 1900 and the mean of 1946. Moreover, Table 3 demonstrates the buckets and percentages of these census tract characteristics and the distribution of these buckets in different PDC.

Fig. 4a shows the distribution of energy burden in Chicago at each census tract level. In the United States, energy poverty is frequently identified by households whose energy burden surpasses 6 % of their household income. This definition is based on the guideline that energy expenditures should not exceed 20 % of housing costs, and housing costs themselves should not surpass 30 % of household income [66,75]. As illustrated in Fig. 4a, over 53.99 % census tract has an energy burden between 1 and 2%, followed 37.68 % census track with an energy burden of 3–5%, 8.33 % census track with an energy burden of 6–10 %. Fig. 4b shows 8.61 % of census tract has less than 0.18 % with external shading, 27.06 % of census track has 0.18–0.26 % of buildings with external shading, 27.83 % census track has 0.26–0.32 % of buildings with external shading, 23.32 % census track has 0.32–0.39 % buildings with external shading, and 13.19 % census track has 0.39–0.63 % buildings with external shading.

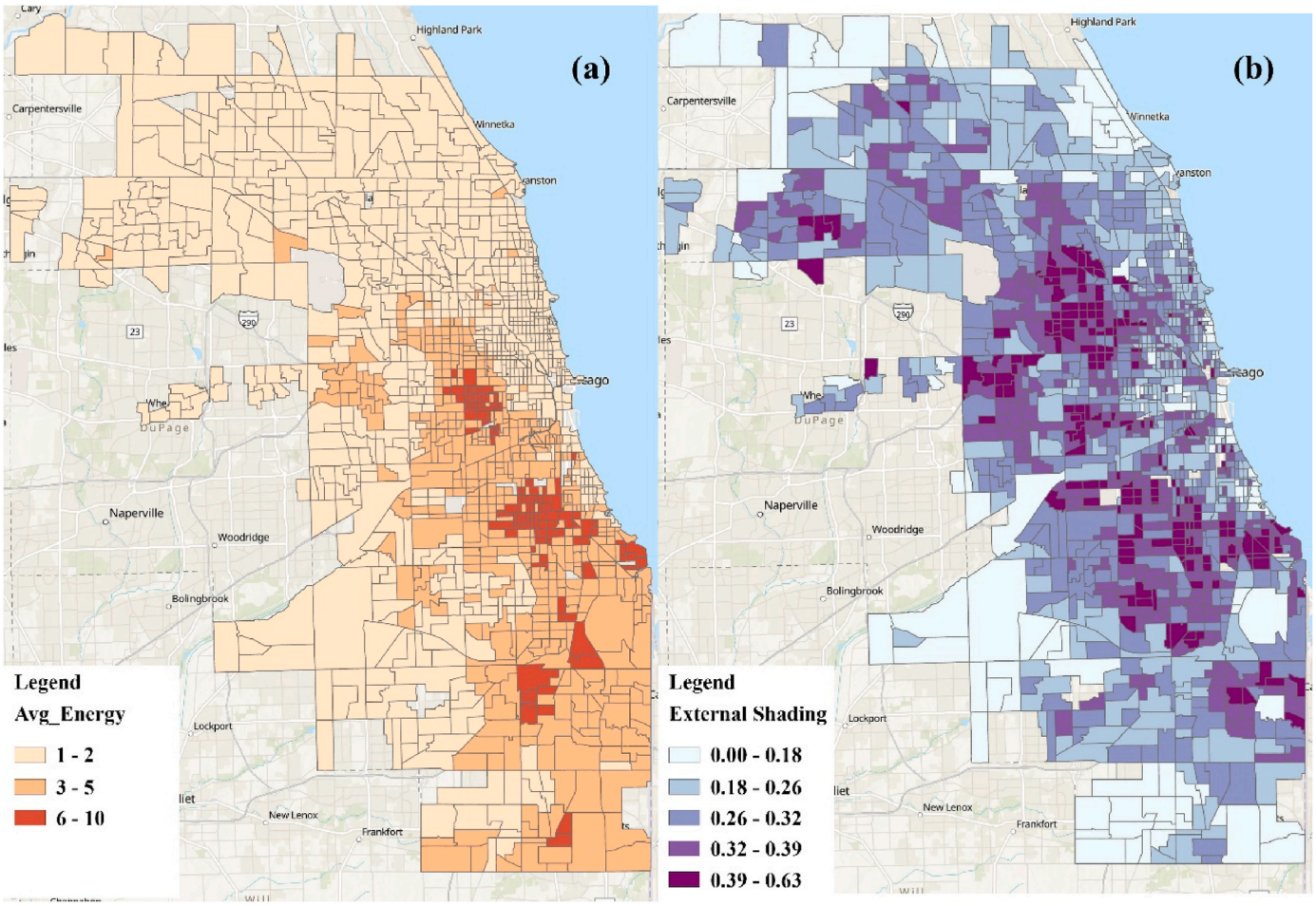


Fig. 4. Tertile maps of Chicago at census tract level (a) the average energy burden (b) external shading.

Fig. 5 shows the distribution of WWR in Chicago. When examined together with Fig. 4a, In general, we observe the positive relation between energy burden and external shading, WWR-Fixed, and negative relation between energy burden to WWR-Nonfixed. Such correlation is further tested in statistical analysis.

4.2. Statistical results

Table 4 displays the multiple linear regression model results for 3 independent variables of WWR-Fixed, WWR-Nonfixed, and external shading. Prob (F-Statistic) provides an indication of the acceptability of the hypothesis, specifically whether the assumption that the effects of the variables under consideration are zero holds true. In the present scenario, the Prob (F-Statistic) is less than 0.005, which indicates that the overall regression model is statistically significant, suggesting that the independent variables, as a group, have a significant effect on the energy burden. Therefore, it gives an initial interpretation that the null hypothesis is rejected, the alternative hypothesis is accepted which is the Energy burden is associated with WWR and external shading. The R-squared result shows a value of 0.077 which means that the current regression model explains only 7.7 % of the variances of the Energy Burden based on the current independent variables. Then looking into individual independent variables, all three independent variables have p-values less than 0.005, indicating that they are statistically significant predictors of the energy burden. The model equation based on the regression analysis findings may be expressed as.

The coefficients (−17.7497, −62.4676, and 7.5784) represent the estimated change in the average energy burden for each unit increase in the corresponding independent variable. The negative coefficient for WWR-Fixed suggests that an increase in this variable is associated with a decrease in the energy burden. Similarly, an increase in the external shading is associated with an increase in the energy burden, as indicated by the positive coefficient for external shading variable. The potential interpretation for a negative correlation between WWR-Fixed to energy burden is related to the high heating energy demand in the wintertime in Chicago areas. Since the Chicago area is in a heating-dominating climate zone, larger window areas are beneficial for direct solar heat gain with sufficient insulation. Previous studies showed a range of optimum WWR of 30–45 % was found optimal in most climate zones, which is higher than the average in studied census tract areas [76]. Another study showed quintuple windows with a WWR of 60 % provided the best energy performance in Estonia [77]. Along the same rationale, the positive correlation between the existence of external shadings and increased energy burden can be explained, the more shades, the less direct solar heat gain, therefore increased heating demand. The negative correlation between the WWR-Nonfixed to the energy burden can be explained by the potential of natural ventilation, which can cool down the building naturally without relying on air-conditioning.

Although this model showed the significance of the association of the PDC to the average energy burden, it could not explain all the complexity and variances of the model due to the low R^2 metric. After reviewing and verifying these results, other regression methods were tested to investigate whether there was any better method to explain the dependent variable. However, none of the additional regression methods were able to explain the complexity of the model and produced high errors as well (see Appendix .1). Therefore, this model proved the association of PDC and average energy burden but yet it is not a good fit for predicting this value. To address this issue and test another

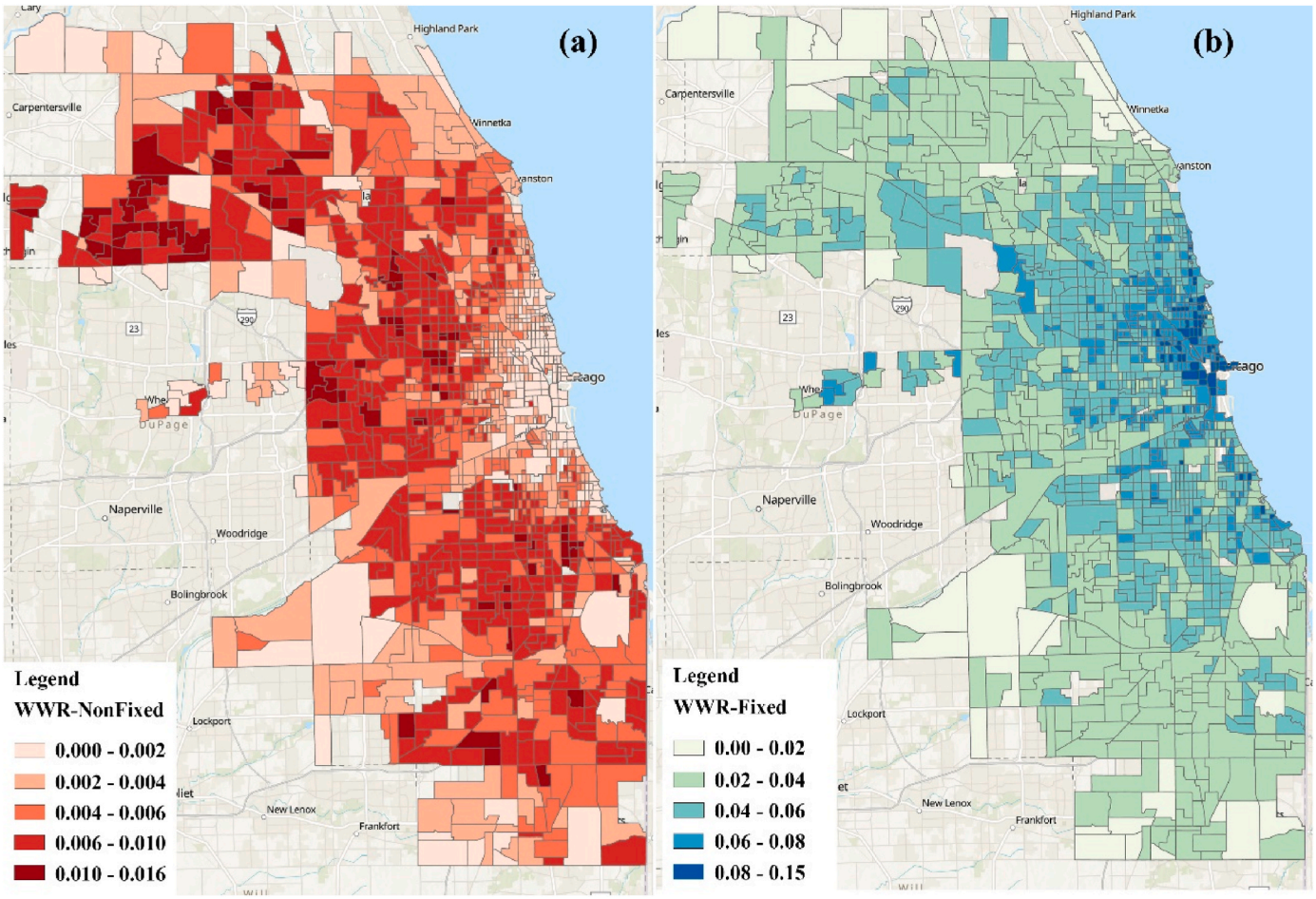


Fig. 5. WWR heatmap. (a) The WWR-Nonfixed denotes the window-to-wall ratio of non-fixed windows (b) the WWR-Fixed indicates the fixed window-to-wall ratio.

Table 2
Descriptive statistics of building characteristics and energy burden at census tract level.

Variables	N (census tract)	Mean (SD)	Std. dev.	Min	Max
Passive design variables	1441				
Window-to-Wall ratio (fixed)		0.0482	0.0180	0	0.1420
Window-to-Wall ratio (non-fixed)		0.0053	0.0029	0	0.0158
External shading		0.0292	0.083	0	
Energy burden	1441				
Energy burden		2.812	1.674	1	10

hypothesis regarding the impact of demographic variables on average energy burden, more covariates were added to the model to improve the model accuracy and reduce the errors. The percentage of the total population between 65 and 74 years old, the percentage of the total female population, the population of Asian people, the population of Black or African American people, the in-poverty population, median year of building structure, the population of Hispanics, the population of over 65-year-old people were added as the covariates to achieve this. Table 5 demonstrates the summary of these data.

Pearson and Kendall correlation was used to prepare an insight about the relationship between the different variables. The results of the

Table 4
Statistical Analysis Results (OLS regression model).

Prob > F = 0.0000, F-Statistic = 39.91
Multiple R-squared = 0.077
Adj Multiple R-squared = 0.075

Variables	Coefficients	Std err	*t* value	*p* value
Intercept	1.7854	0.169	10.573	0.0000
WWR-Fixed	−17.749	3.212	−5.525	0.0000
WWR-Nonfixed	−62.467	21.701	−2.878	0.00405
External shading	7.478	0.819	9.335	0.0000

WWR-Fixed = Fixed Window-toWall Ratio; WWR-Nonfixed = Non-Fixed Window-to-Wall Ratio.

Table 3
Buckets and Percentages of neighborhood characteristics and energy burden, census tract.

Buckets	1	2	3.	4	5
Window-to-Wall ratio (fixed)	<0.02 % (3.12 %)	0.02–0.04 % (31.92 %)	0.04–0.06 % (42.54 %)	0.06–0.08 % (17.63 %)	0.08–0.15 % (4.79 %)
Window-to-Wall ratio (non-fixed)	<0.002 % (13.60 %)	0.002–0.004 % (21.79 %)	0.004–0.006 % (25.05 %)	0.006–0.010 % (33.24 %)	0.010%–0.016 % (6.32 %)
External shading	<0.18 % (8.61 %)	0.18–0.26 % (27.06 %)	0.26–0.32 % (27.83 %)	0.32–0.39 % (23.32 %)	0.39–0.63 % (13.19 %)
Energy burden	1–2% (53.99 %)	3–5% (37.68 %)	6–10 % (8.33 %)		

* Buckets represent the range each variable fits into and percentages represent the distribution of these buckets.

Table 5
Summary of the dependent variables, independent variables, and covariates.

Variables	N (census tract)	mean	std	min	25 %	50 %	75 %	max
Avg_Energyburden	1402	2.800	1.663	1	2	2	4	10
WWR-Fixed	1402	0.048	0.018	0.001	0.036	0.047	0.058	0.142
WWR-Nonfixed	1402	0.005	0.003	0	0.003	0.005	0.007	0.016
External_shade	1402	0.293	0.083	0.012	0.238	0.292	0.348	0.622
% Total Population: 65–74 Years	1402	8.502	3.786	0	6.040	8.110	10.638	31.780
% Total Population: Female	1402	51.543	4.886	0.48	48.800	51.260	54.048	100
% Total Population: Black or African American Alone	1402	26.180	35.019	0	1.703	5.725	48.293	99.92
% Total Population: American Indian and Alaska Native Alone	1402	0.285	0.716	0	0.000	0.000	0.220	7.090
% Total Population: Asian Alone	1402	6.311	9.411	0	0.263	2.710	8.360	84.910
Population for Whom Poverty Status Is Determined	1402	3811.633	1670.699	8	2504	3727	4945	9812
Median Year Structure Built	1400	1914.382	249.358	1900	1900	1955	1968	2008
pctpop_over65y	1402	14.857	6.922	0	10.120	14.280	18.708	54.980
pctpop_hispanic	1402	22.633	25.634	0	4.443	11.645	30.065	99.120

Avg_Energyburden = Average Energy Burden; WWR-Fixed = Fixed Window-toWall Ratio; WWR-Nonfixed = Non-Fixed Window-to-Wall Ratio; External_shade = External Shading; Population for Whom Poverty Status Is Determined = Population of the people in poverty; pctpop_over65y = percentage of the population over 65 years; pctpop_hispanic = percentage of the Hispanic population.

Pearson correlation are shown in Fig. 6. The results showed that a significant positive correlation with the "Avg_EnergyBurden" can be observed in "Total population of Black or African American alone" (0.62), a moderate correlation can be in "Total population of Asian Alone" (−0.38), and a weak correlation in "Population with Poverty Status" (−0.27), and "External shading" (0.24). It is worth mentioning that there are potentially variable interpretations for those positive and negative correlations. For example, the negative correlations between energy burden and the variables "Asian alone population" and "Poverty status" may have negative correlations with the average energy burden, which can be derived from different causes. The fact that higher poverty status is correlated with lower energy burden doesn't necessarily mean that households in poverty experience less energy burden. Instead, these households might spend the majority of their income on food, rent, and other essential expenses, and the trade-off is a less comfortable living condition without heating or cooling. A deep dive into further research can provide clarity and explanation.

The second OLS regression was implemented to investigate the new variables' results on the model. As shown in Table 6, the second model's R-squared value has increased (from 0.077 in Table 3 to 0.560), denoting that the additional demographic variables helped better explain the variation of energy burden in different census tracts. The second model's F-statistic sharply increased, demonstrating the overall statistical significance of the model. This highlights the general improvement in model fit, which is probably caused by the greater number of variables. However, in the second model, WWR-Fixed and WWR-Nonfixed coefficients both underwent significant changes, but their p-values dramatically increased, indicating lower statistical significance. This might result from multicollinearity or the addition of other variables that explain similar variation.

To verify the multicollinearity, VIF was calculated (>10 indicate significant multicollinearity) to determine that whether the predictor variable should be investigated further or potentially removed from the model. According to the results of the VIF test (See Appendix 3.), all VIF values are less than 10. Moreover, tolerance values are all above 0.2, with the lowest being around 0.25, which suggests that multicollinearity is not severe. Therefore, the increase of the P-values in the PDC shows that there is a mismatch between the GSV data and the demographic data in the linear regression method. Moreover, the low R^2 value (0.56) indicates that the model can only explain 56 % of the variation in energy burden. Consequently, ML-based regressions were employed using a grid search fine-tuning algorithm to find a more suitable model for predicting the average energy burden. To accomplish this, Decision Tree Regression (DTR), Random Forest Regression (RFR), and Support Vector Regression (SVR) were used. The complete set of parameters for grid search can be seen in Appendix 4.

The results of Table 7 show that the SVR method produces the highest accuracy and lowest errors among all the models, with 74.2 % accuracy. This analysis also demonstrates that the inclusion of covariates greatly aids in producing improved models and reducing errors, indicating an association with the average energy burden. Additionally, it highlights that the radial basis function serves as a superior regression method, whereas the linear method falls short of generating satisfactory models. Following SVR, the random forest regression achieves the highest accuracy with 70.5 % accuracy. Overall, a comparison between the results of the ML models prior to adding the covariates (see Appendix 1) and Table 7 allows us to conclude that while the PDC are associated with the average energy burden, the complexity and variations of the dependent variables demand the incorporation of more demographic information and more intricate mathematical models beyond linear regression. Furthermore, the ML-based regression models have proved to be significantly valuable in addressing this issue.

5. Discussion

5.1. Methodology contribution policy implication

In this research, we showcased the viability and precision of machine learning-assisted computer vision models for autonomously identifying and describing passive design attributes in housing. Furthermore, as we applied our algorithms to diverse housing styles during our initial trials, we observed that the models exhibited overall applicability to different building types. The proposed methodology holds promise for broader applicability, as it can be employed in diverse cities and regions to investigate energy-related concerns. Furthermore, the resulting comprehensive database for the Chicago area holds value not only for researchers specializing in the built environment but also for public health researchers, policymakers, and potential developers seeking to make informed decisions pertaining to energy usage and investments.

By analyzing a dataset of 227,000 GSV images, our study delved into the correlations between passive design indicators and household energy burden. This research introduces a distinctive methodology to the growing body of studies investigating the potential influence of housing attributes on energy burden. Our study found that a higher window-to-wall ratio and less external shading was associated with lower energy burden. Overall, a larger window area, as compared to the current average of 4.8 %, is beneficial for reducing the energy demand both in winter and summer. This provides empirical evidence for future building renovation activities; such findings are in line with previous research on housing renovation. The study found by improving housing conditions, the estimated low-income household energy burden can be reduced by 25 %.[1] More empirical evidence will help decisionmakers to make informative policies that adopt and promote low-cost passive design to elevate the energy burden for low-income families. Moreover, it was

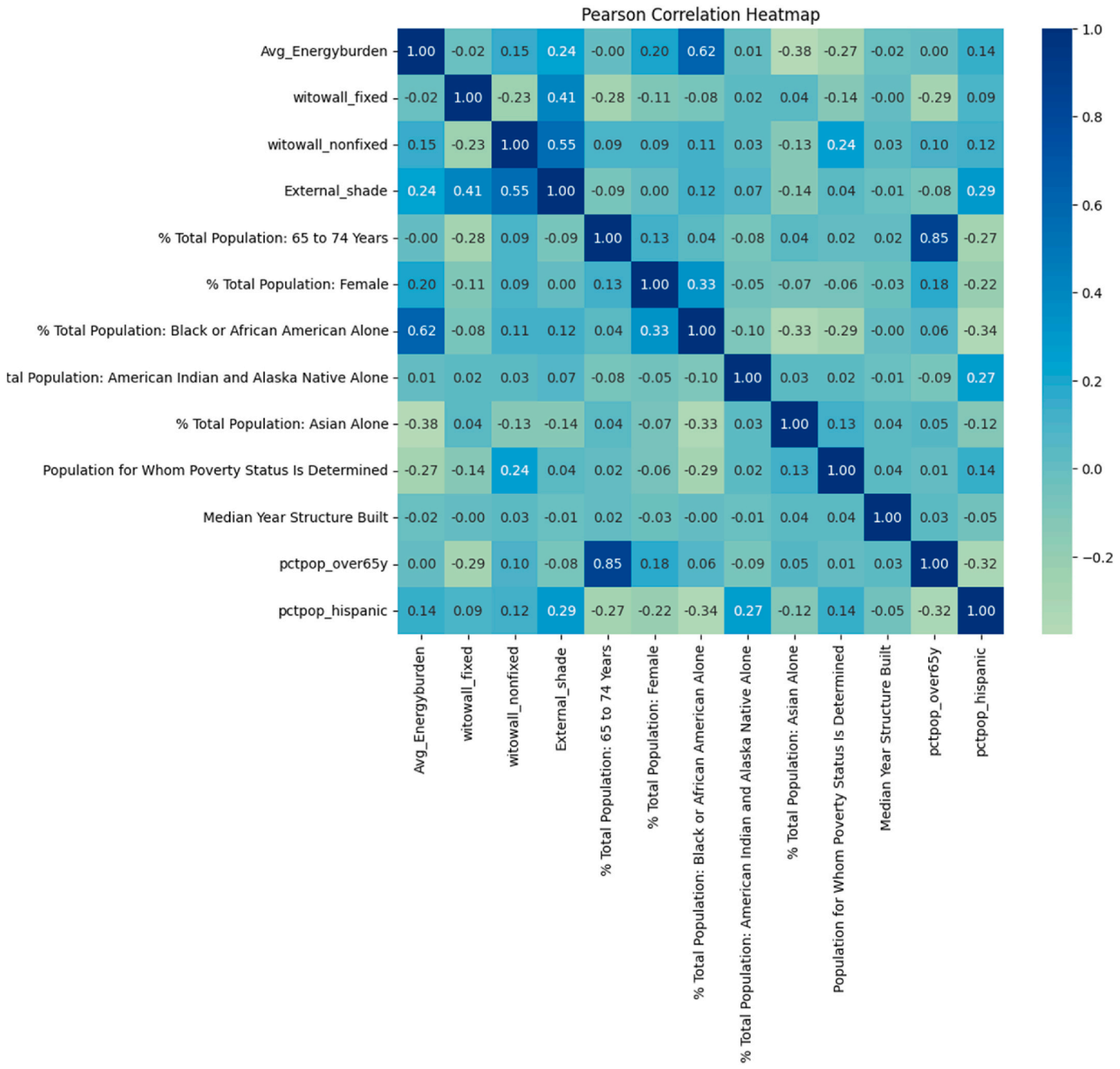


Fig. 6. Pearson correlation heatmap results.

shown that although building characteristics are associated with average energy burden, they are not enough to observe the behavior of this variable, and the researchers and policymakers need to pay attention to demographic variables as well. Another finding was that the linear models lack the capability of capturing the average energy burden variances, and more comprehensive ML-based algorithms, specifically SVR with RBF method, are needed to cover this complexity.

5.2. Study finding in context

Previous research has leveraged Google Street View (GSV) for virtual assessments related to neighborhood walkability [78], physical disorder [79], and urban greenery at the neighborhood level. Furthermore, studies comparing on-site audits to virtual assessments have shown consistent findings across various metrics [80]. However, the application of GSV for building audits has been somewhat restricted, and to the best of our knowledge, no study has specifically concentrated on passive design features, likely due to the widespread reliance on air conditioning in the United States.

Air conditioning is a prevalent feature in American households, with approximately 90 % of U.S. households reported to have some form of air conditioning system, as indicated by data from the 2015 Residential Energy Consumption Survey [81]. In contrast, fewer than 10 % of European households possess air conditioning [82]. Furthermore, the adoption of air conditioning in developing nations such as China and India are experiencing rapid expansion. The effects of climate change are evident in the increasing prevalence of extreme heat and cold events worldwide, as exemplified by the 2022 heat wave [83]. These recurrent heatwaves are expected to further boost the demand for and reliance on air conditioning systems. The global sales of air conditioning systems

Table 6
The OLS Regression results on the combination of dependent variables and covariates.

Prob > F = 0.0000, F-Statistic = 160.8				
Multiple R-Squared = 0.560				
Adj Multiple R-Squared = 0.556				
Variables	Coefficients	Std err	*t* value	*p* value
Intercept	0.5308	0.441	1.205	0.228
WWR-Fixed	1.9438	2.414	0.805	0.421
WWR-Nonfixed	12.5332	15.641	0.801	0.423
External_shade	0.451	0.617	0.731	0.465
% Total Population: 65–74 Years	0.009	0.015	0.602	0.547
% Total Population: Female	0.0112	0.007	1.712	0.087
% Total Population: Asian Alone	−0.0142	0.003	−4.095	0
% Total Population: Black or African American Alone	0.325	0.001	29.474	0
Population for Whom Poverty Status Is Determined	−0.0001	1.96E-05	−5.903	0
Median Year Structure Built	6.18E-05	0	0.518	0.605
pctpop_hispanic	0.0264	0.001	17.982	0
pctpop_over65y	0.0192	0.008	2.264	0.024

Avg_Energyburden = Average Energy Burden; WWR-Fixed = Fixed Window-toWall Ratio; WWR-Nonfixed = Non-Fixed Window-to-Wall Ratio; External_shade = External Shading; Population for Whom Poverty Status Is Determined = Population of the people in poverty; pctpop_over65y = percentage of the population over 65 years; pctpop_hispanic = percentage of the Hispanic population.

more than tripled between 1990 and 2016, and this growth trend is anticipated to persist [82].

There are three significant challenges associated with excessive reliance on air conditioning systems. Firstly, it leads to increased energy consumption, with heating and cooling accounting for approximately 51.6 % of total energy usage in residential buildings [84]. Secondly, it places strain on the power grid and elevates the risk of power outages, primarily due to the heightened energy demand during heatwaves, especially in developing nations where the original power grid capacity is stretched to its limits. The third challenge pertains to environmental inequality. The escalating impact of extreme weather events driven by climate change, coupled with disparities in disaster response systems across countries, disproportionately affects low-income households and vulnerable communities.

For instance, heatwaves are the leading cause of fatalities in numerous countries, including developed nations like the United States and Australia, as well as in cooler climate regions such as Northern Europe. Fatality statistics predominantly encompass individuals from vulnerable groups, such as the elderly (65 years and older), young children, pregnant women, people with underlying health conditions like diabetes, obesity, hypertension, respiratory ailments, or mental health issues, as well as those from low-income households or facing social isolation [85,86]. To address the disparities in both energy consumption and public health, passive design strategies should be seriously considered and integrated as a central component.

5.3. Study strength and limitation

The housing condition and quality can impact health and impose an energy burden on residents. The aging housing stock in the United States is characterized by deteriorating conditions, which have given rise to significant environmental and health issues, including mold infestation, lead contamination, and indoor heat exposure. Up to this point, the process of characterizing and examining housing conditions and passive design attributes has predominantly been confined to individual building studies. This limitation stems from the resource-intensive demands associated with (1) conducting on-site inspections and energy audits to evaluate building conditions and (2) manually annotating street images. In this study, we proposed, tested, and validated a DL-enabled computer vision model and ML-enhanced regression model to study city-wide study at census tract level. This novel approach tapped into the massive sources of street image data that have not been fully utilized in built environment research. The method and data produced can be used by other built environmental researchers.

This study also presents several limitations within its research scope. Similar to other data collection methods, image data can only encompass a portion of the building features that influence energy demand and energy burden. For instance, it is not feasible to establish indicators for deteriorating exterior facades using this approach. Another constraint arises from the types of building features that can be readily extracted through computer vision models. Additionally, the depth of detail in extracting building features may be restricted due to the emerging nature of the database. The creation of a deep learning-powered computer vision model capable of accurately extracting each of these features would necessitate a substantial training dataset encompassing various conditions. The second limitation of the study was that the data and analyses were only from one geographic region. More cross-regional studies would be beneficial to validate the extracted passive design features. The third limitation is that the analyses did not consider other housing characteristics, especially inside features such as heating and cooling systems and insulation. Moreover, the concentration of this research was on passive design variables and did not consider the external factors such as adjacent building's shading impact which can be the topic of another study. Even though, this research achieved relatively high accuracy in predicting the average energy burden based on passive design and demographic data, this research can act as a benchmark for future research to examine other sources of information and variables to assess the possibility of reaching even better results.

6. Conclusion

The challenges of characterizing housing conditions and features across extensive geographical areas pose a barrier to examining the links between passive design and energy burden on a national level. In an effort to enhance research in this field and contribute insights for mitigating energy burden, we initiated a project utilizing 277,000 GSV images and harnessing the capabilities of computer vision models. This paper introduces a completely automated and scalable workflow designed for the detection of three passive design features in all buildings across the Chicago metropolitan area. This methodology relies on geotagged street view imagery data and leverages this information to create a predictive platform for average energy burden.

According to a recent study, over one-third of American households experience a higher energy burden [66]. Despite its prevalence as a

Table 7
The results of the SVR, DTR, and RFR ML methods.

Algorithm	Parameters	R-Squared Value	MAE	MSE	RMSE
Decision Tree	max_depth = 5, max_features = 'log2', min_samples_leaf = 4, min_samples_split = 5	0.471	0.933	1.681	1.297
Random Forest Regression	max_depth = 80, max_features = 3, min_samples_leaf = 3, min_samples_split = 8, n_estimators = 200	0.705	0.689	0.937	0.968
SVR	Default Parameters – Kernel: Linear	0.597	0.790	1.281	1.132
SVR	(C = 10, epsilon = 0.1, gamma = 0.1, kernel = 'rbf')	0.676	0.717	1.029	1.014
SVR[a]	(C = 100, epsilon = 0.1, gamma = 0.01, kernel = 'rbf')	0.742	0.663	0.818	0.904

[a] The star mark denotes the best model among all the models.

social concern in the United States, little is known about the relationship between energy burden and the mitigation benefits of low-cost passive design features. This is due to the lack of passive design data at large scales, as traditional field audits are time-consuming and labor-intensive. By employing an innovative approach that encompasses computer vision techniques and ML algorithms, this project seeks to contribute to the advancement of knowledge in the field of energy efficiency and sustainable design. The results and insights derived from this study indicate that this combined approach using GSV images has the potential to inform policy decisions, facilitate evidence-based urban planning, and mitigate energy burden through appropriate housing renovation.

The findings from this study also indicate that more housing condition variables (e.g., insulation) and passive design features (e.g., cool roof) should be studied using the proposed method and approach to achieve more robust results. The next steps for the research team are (a) to include more passive design variables, (b) to test the computer vision model and algorithm in other geographic areas, and (c) to include more social, economic, and demographic factors in the study.

Funding

Research reported in this publication was supported by the Lucy Institute of Data and Society, the University of Notre Dame, the National Library of Medicine, and the National Institute on Minority Health and Health Disparities under Award Numbers R01LM012849 and R01MD016037 (Q.C.N.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Notre Dame and National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

CRediT authorship contribution statement

Siavash Ghorbany: Writing – original draft, Investigation, Formal analysis. **Ming Hu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Siyuan Yao:** Visualization. **Chaoli Wang:** Writing – review & editing, Visualization. **Quynh Camthi Nguyen:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Xiaohe Yue:** Validation, Investigation, Formal analysis, Data curation. **Mitra Alirezaei:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Tolga Tasdizen:** Writing – review & editing, Validation, Supervision, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Matthew Sisk:** Software, Resources, Data curation.

Declaration of competing interest

The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2023.111126.

References

[1] Department of Energy, Low-Income Community Energy Solutions, Energy.Gov., 2023. https://www.energy.gov/scep/slsc/low-income-community-energy-solutions.
[2] A. Drehobl, L. Ross, R. Ayala, How high are household energy burdens, An Assess. Natl. Metrop. Energy Burdens across US. (2020).
[3] K. Zhang, T.-H. Chen, C.E. Begley, Impact of the 2011 heat wave on mortality and emergency department visits in Houston, Texas, Environ. Heal. 14 (2015) 11, https://doi.org/10.1186/1476-069X-14-11.
[4] G.A. Meehl, C. Tebaldi, More intense, more frequent, and longer lasting heat waves in the 21st century, Science 305 (2004) 994–997, https://doi.org/10.1126/science.1098704, 80-.
[5] S. Kheirinejad, O. Bozorg-Haddad, V.G. Gude, The water, food and energy nexus, in: Water Resour. Futur. Perspect. Challenges, Concepts Necessities, IWA Publishing, 2021, pp. 175–204, https://doi.org/10.2166/9781789062144_0175.
[6] Y. Guo, A. Gasparrini, S. Li, F. Sera, A.M. Vicedo-Cabrera, M. de Sousa Zanotti Stagliorio Coelho, P.H.N. Saldiva, E. Lavigne, B. Tawatsupa, K. Punnasiri, Quantifying excess deaths related to heatwaves under climate change scenarios: a multicountry time series modelling study, PLoS Med. 15 (2018) e1002629.
[7] M.G. Kent, N.K. Huynh, A.K. Mishra, F. Tartarini, A. Lipczynska, J. Li, Z. Sultan, E. Goh, G. Karunagaran, A. Natarajan, A. Indrajith, I. Hendri, K.I. Narendra, V. Wu, N. Chin, C.P. Gao, M. Sapar, A. Seoh, N. Shuhadah, S. Valliappan, T. Jukes, C. Spanos, S. Schiavon, Energy savings and thermal comfort in a zero energy office building with fans in Singapore, Build. Environ. 243 (2023) 110674, https://doi.org/10.1016/j.buildenv.2023.110674.
[8] S. Koley, Challenges in sustainable development of smart cities in India, Sustainability 13 (2020) 155–160, https://doi.org/10.1089/sus.2020.0017.
[9] X. Chen, H. Yang, Sensitivity analysis and optimization of a typical passively designed residential building with hybrid ventilation in hot and humid climates, Energy Proc. 142 (2017) 1781–1786, https://doi.org/10.1016/j.egypro.2017.12.563.
[10] D.K. Bhamare, M.K. Rathod, J. Banerjee, Passive cooling techniques for building and their applicability in different climatic zones—the state of art, Energy Build. 198 (2019) 467–490, https://doi.org/10.1016/j.enbuild.2019.06.023.
[11] H. Campaniço, P. Hollmuller, P.M.M. Soares, Assessing energy savings in cooling demand of buildings using passive cooling systems based on ventilation, Appl. Energy 134 (2014) 426–438, https://doi.org/10.1016/j.apenergy.2014.08.053.
[12] M. Salameh, B. Touqan, Traditional passive design solutions as a key factor for sustainable modern urban designs in the hot, arid climate of the United Arab Emirates, Buildings 12 (2022) 1811, https://doi.org/10.3390/buildings12111811.
[13] M. Zune, L. Rodrigues, M. Gillott, Vernacular passive design in Myanmar housing for thermal comfort, Sustain. Cities Soc. 54 (2020) 101992, https://doi.org/10.1016/j.scs.2019.101992.
[14] S. Srivastav, P.J. Jones, Use of traditional passive strategies to reduce the energy use and carbon emissions in modern dwellings, Int. J. Low Carbon Technol. 4 (2009) 141–149, https://doi.org/10.1093/ijlct/ctp021.
[15] M. Hu, K. Zhang, Q. Nguyen, T. Tasdizen, The effects of passive design on indoor thermal comfort and energy savings for residential buildings in hot climates: a systematic review, Urban Clim. 49 (2023) 101466, https://doi.org/10.1016/j.uclim.2023.101466.
[16] C. Basu, V.K. Paul, M.M. Syal, Performance indicators for energy efficiency retrofitting in multifamily residential buildings, J. Green Build. 14 (2019) 109–136.
[17] C.K. Cheung, R.J. Fuller, M.B. Luther, Energy-efficient envelope design for high-rise apartments, Energy Build. 37 (2005) 37–48, https://doi.org/10.1016/j.enbuild.2004.05.002.
[18] R. Han, Z. Xu, Y. Qing, Study of passive evaporative cooling technique on water-retaining roof brick, Procedia Eng. 180 (2017) 986–992, https://doi.org/10.1016/j.proeng.2017.04.258.
[19] R. Yao, V. Costanzo, X. Li, Q. Zhang, B. Li, The effect of passive measures on thermal comfort and energy conservation. A case study of the hot summer and cold winter climate in the Yangtze River region, J. Build. Eng. 15 (2018) 298–310, https://doi.org/10.1016/j.jobe.2017.11.012.
[20] N.C. Onyenokporo, E.T. Ochedi, Low-cost retrofit packages for residential buildings in hot-humid Lagos, Nigeria, Int. J. Build. Pathol. Adapt. 37 (2019) 250–272, https://doi.org/10.1108/IJBPA-01-2018-0010.
[21] A.P. Vargas, L. Hamui, Thermal energy performance simulation of a residential building retrofitted with passive design strategies: a case study in Mexico, Sustainability 13 (2021) 8064, https://doi.org/10.3390/su13148064.
[22] IBM, What Is Machine Learning? | IBM, Int. Bus. Mach. Corp., 2023. https://www.ibm.com/topics/machine-learning.
[23] I. El Naqa, M.J. Murphy, What is machine learning? in: I. El Naqa, R. Li, M. J. Murphy (Eds.), Mach. Learn. Radiat. Oncol. Springer International Publishing, Cham, 2015, pp. 3–11, https://doi.org/10.1007/978-3-319-18305-3_1.
[24] R.P. Masini, M.C. Medeiros, E.F. Mendes, Machine learning advances for time series forecasting, J. Econ. Surv. (2023) 76–111, https://doi.org/10.1111/joes.12429.
[25] L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, S. Salcedo-Sanz, Machine learning regressors for solar radiation estimation from satellite data, Sol. Energy 183 (2019) 768–775, https://doi.org/10.1016/j.solener.2019.03.079.
[26] D.J. Silva, J. Ventura, J.P. Araújo, Predicting the performance of magnetocaloric systems using machine learning regressors, Energy AI 2 (2020) 100030, https://doi.org/10.1016/j.egyai.2020.100030.
[27] S. Ghorbany, E. Noorzai, S. Yousefi, BIM-based solution to enhance the performance of public-private partnership construction projects using copula

bayesian network, Expert Syst. Appl. 216 (2023) 119501, https://doi.org/10.1016/j.eswa.2023.119501.
[28] S.T.S. Bukkapatnam, K. Afrin, D. Dave, S.R.T. Kumara, Machine learning and AI for long-term fault prognosis in complex manufacturing systems, CIRP Ann. 68 (2019) 459–462, https://doi.org/10.1016/j.cirp.2019.04.104.
[29] G. Rebala, A. Ravi, S. Churiwala, Machine Learning Definition and Basics BT - an Introduction to Machine Learning, 2019, https://doi.org/10.1007/978-3-030-15729-6_1.
[30] V. Kathiresan, S. Karthik, D. Prabakar, M.S. Kavitha, Logistic Regression-Based Machine Learning Model for Mutation Classification in the Discovery of Precision Medicine, EAI/Springer Innov. Commun. Comput., 2023, pp. 81–92, https://doi.org/10.1007/978-3-031-27700-9_6.
[31] S. Rakshit, N. Clement, N.R. Vajjhala, Exploratory review of applications of machine learning in finance sector, in: Lect. Notes Data Eng. Commun. Technol., 2022, pp. 119–125, https://doi.org/10.1007/978-981-16-5685-9_12.
[32] T. Siddiqui, A.Y.A. Amer, A comprehensive review on text classification and text mining techniques using spam dataset detection, Math. Comput. Sci. 2 (2023) 1–17, https://doi.org/10.1002/9781119896715.ch1. Wiley.
[33] Z. Kearl, J. Vogel, Urban extreme heat, climate change, and saving lives: lessons from Washington state, Urban Clim. 47 (2023) 101392, https://doi.org/10.1016/j.uclim.2022.101392.
[34] U. Grömping, Variable importance in regression models, Wiley Interdiscip. Rev. Comput. Stat. 7 (2015) 137–152, https://doi.org/10.1002/wics.1346.
[35] H. Luo, F. Cheng, H. Yu, Y. Yi, SDTR: soft decision tree regressor for tabular data, IEEE Access 9 (2021) 55999–56011, https://doi.org/10.1109/ACCESS.2021.3070575.
[36] S. Kim, M. Jeong, B.C. Ko, Self-supervised keypoint detection based on multi-layer random forest regressor, IEEE Access 9 (2021) 40850–40859, https://doi.org/10.1109/ACCESS.2021.3065022.
[37] R. Khemchandani, K. Goyal, S. Chandra, Generalized eigenvalue proximal support vector regressor for the simultaneous learning of a function and its derivatives, Int. J. Mach. Learn. Cybern. 9 (2018) 2059–2070, https://doi.org/10.1007/s13042-017-0687-3.
[38] H. Yasin, R.E. Caraka, A. Hoyyi, Prediction of crude oil prices using support vector regression (SVR) with grid search - cross validation algorithm, Global J. Pure Appl. Math. 12 (2016) 3009–3020. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84980360979&partnerID=40&md5=9f13e14d32dd5ed63d2e0a7a1b13de74.
[39] S.N.V.B. Rao, V.P.K. Yellapragada, K. Padma, D.J. Pradeep, C.P. Reddy, M. Amir, S. S. Refaat, Day-ahead load demand forecasting in urban community cluster microgrids using machine learning methods, Energies 15 (2022), https://doi.org/10.3390/en15176124.
[40] G. Ben Brahim, Weather conditions impact on electricity consumption in smart homes: machine learning based prediction model, in: 2021 8th Int. Conf. Electr. Electron. Eng., IEEE, 2021, pp. 93–98, https://doi.org/10.1109/ICEEE52452.2021.9415917.
[41] B. Yarahmadi, S.M. Hashemianzadeh, S.M.-R. Milani Hosseini, Machine-learning-based predictions of imprinting quality using ensemble and non-linear regression algorithms, Sci. Rep. 13 (2023) 12111, https://doi.org/10.1038/s41598-023-39374-1.
[42] P.G.S. Mohith, P. Madhava Krishna, A. Velmurugan, Energy price forecasting in Python using machine learning algorithm, in: Lect. Notes Electr. Eng., 2021, pp. 621–631, https://doi.org/10.1007/978-981-15-8752-8_63.
[43] S. Namany, T. Al-Ansari, Energy, water, food nexus decision-making for sustainable food security, in: Environ. Footprints Eco-Design Prod. Process., 2021, pp. 191–216, https://doi.org/10.1007/978-981-16-0239-9_7.
[44] F. Zhang, C. Deb, S.E. Lee, J. Yang, K.W. Shah, Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique, Energy Build. 126 (2016) 94–103, https://doi.org/10.1016/j.enbuild.2016.05.028.
[45] M. Kent, N.K. Huynh, S. Schiavon, S. Selkowitz, Using support vector machine to detect desk illuminance sensor blockage for closed-loop daylight harvesting, Energy Build. 274 (2022) 112443, https://doi.org/10.1016/j.enbuild.2022.112443.
[46] S. Zhu, C. Ma, Y. Zhang, K. Xiang, A hybrid metamodel-based method for quick energy prediction in the early design stage, J. Clean. Prod. 320 (2021) 128825, https://doi.org/10.1016/j.jclepro.2021.128825.
[47] J.-F. Toubeau, L. Pardoen, L. Hubert, N. Marenne, J. Sprooten, Z. De Grève, F. Vallée, Machine learning-assisted outage planning for maintenance activities in power systems with renewables, Energy 238 (2022) 121993, https://doi.org/10.1016/j.energy.2021.121993.
[48] A. Buylova, Spotlight on energy efficiency in Oregon: investigating dynamics between energy use and socio-demographic characteristics in spatial modeling of residential energy consumption, Energy Pol. 140 (2020) 111439, https://doi.org/10.1016/j.enpol.2020.111439.
[49] C.E. Kontokosta, V.J. Reina, B. Bonczak, Energy cost burdens for low-income and minority households, J. Am. Plann. Assoc. 86 (2020) 89–105, https://doi.org/10.1080/01944363.2019.1647446.
[50] O. Ma, K. Laymon, M.H. Day, R. Bracho, J.D. Weers, A.J. Vimont, Low-Income Energy Affordability Data (LEAD) Tool Methodology, 2019.
[51] C. Marino, A. Nucara, M. Pietrafesa, Does window-to-wall ratio have a significant effect on the energy consumption of buildings? A parametric analysis in Italian climate conditions, J. Build. Eng. 13 (2017) 169–183, https://doi.org/10.1016/j.jobe.2017.08.001.
[52] Q. Al-Yasiri, M. Alktranee, M. Szabó, M. Arıcı, Building envelope-enhanced phase change material and night ventilation: effect of window orientation and window-to-wall ratio on indoor temperature, Renew. Energy 218 (2023) 119263, https://doi.org/10.1016/j.renene.2023.119263.
[53] L. Troup, R. Phillips, M.J. Eckelman, D. Fannon, Effect of window-to-wall ratio on measured energy consumption in US office buildings, Energy Build. 203 (2019) 109434, https://doi.org/10.1016/j.enbuild.2019.109434.
[54] A. Zhang, R. Bokel, A. Van Den Dobbelsteen, Y. Sun, Q. Huang, Q. Zhang, The effect of geometry parameters on energy and thermal performance of school buildings in cold climates of China, Sustainability 9 (2017) 1708, https://doi.org/10.3390/su9101708.
[55] Q. Tushar, M.A. Bhuiyan, G. Zhang, Energy simulation and modeling for window system: a comparative study of life cycle assessment and life cycle costing, J. Clean. Prod. 330 (2022) 129936, https://doi.org/10.1016/j.jclepro.2021.129936.
[56] Y.H. Liu, Feature extraction and image recognition with convolutional neural networks, J. Phys. Conf. Ser. 1087 (2018) 062032, https://doi.org/10.1088/1742-6596/1087/6/062032.
[57] Y. Liu, H. Pu, D.-W. Sun, Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices, Trends Food Sci. Technol. 113 (2021) 193–204, https://doi.org/10.1016/j.tifs.2021.04.042.
[58] Y. Li, L. Peng, C. Wu, J. Zhang, Street view imagery (SVI) in the built environment: a theoretical and systematic review, Buildings 12 (2022) 1167, https://doi.org/10.3390/buildings12081167.
[59] Ç.F. Özgenel, A.G. Sorguç, Performance Comparison of Pretrained Convolutional Neural Networks on Crack Detection in Buildings, IAARC Publications, 2018, pp. 1–8.
[60] H. Perez, J.H. Tah, A. Mosavi, Deep learning for detecting building defects using convolutional neural networks, Sensors 19 (2019) 3556.
[61] S. Zou, L. Wang, Detecting individual abandoned houses from google street view: a hierarchical deep learning approach, ISPRS J. Photogrammetry Remote Sens. 175 (2021) 298–310, https://doi.org/10.1016/j.isprsjprs.2021.03.020.
[62] R. Castello, S. Roquette, M. Esguerra, A. Guerra, J.-L. Scartezzini, Deep learning in the built environment: automatic detection of rooftop solar panels using Convolutional Neural Networks, J. Phys. Conf. Ser. 1343 (2019) 012034, https://doi.org/10.1088/1742-6596/1343/1/012034.
[63] C. Hentschel, T.P. Wiradarma, H. Sack, Fine Tuning CNNS with Scarce Training Data—Adapting ImageNet to Art Epoch Classification, IEEE, 2016, pp. 3693–3697.
[64] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90, https://doi.org/10.1145/3065386.
[65] X. Yue, A. Antonietti, M. Alirezaei, T. Tasdizen, D. Li, L. Nguyen, H. Mane, A. Sun, M. Hu, R.T. Whitaker, Q.C. Nguyen, Using convolutional neural networks to derive neighborhood built environments from google street view images and examine their associations with health outcomes, Int. J. Environ. Res. Publ. Health 19 (2022), https://doi.org/10.3390/ijerph191912095.
[66] E. Scheier, N. Kittner, A measurement strategy to address disparities across household energy burdens, Nat. Commun. 13 (2022) 288, https://doi.org/10.1038/s41467-021-27673-y.
[67] M. Brusco, Logistic regression via excel spreadsheets: mechanics, model selection, and relative predictor importance, Inf. Trans. Educ. 23 (2022) 1–11, https://doi.org/10.1287/ited.2021.0263.
[68] S. Ghorbany, S. Yousefi, E. Noorzai, Evaluating and optimizing performance of public–private partnership projects using copula Bayesian network, Eng. Construct. Architect. Manag. (2022), https://doi.org/10.1108/ECAM-05-2022-0492.
[69] K. Kottaridi, M. Anna, D. Vasilis, R. Aimilia, N. Vasileios, A regression analysis method for the prediction of olive oil sensory attributes, J. Agric. Food Res. 12 (2023) 100555, https://doi.org/10.1016/j.jafr.2023.100555.
[70] Y. Sun, S. Ding, Z. Zhang, W. Jia, An improved grid search algorithm to optimize SVR for prediction, Soft Comput. 25 (2021) 5633–5644, https://doi.org/10.1007/s00500-020-05560-w.
[71] K. Mei, J. Liu, X. Zhang, N. Rajatheva, J. Wei, Performance analysis on machine learning-based channel estimation, IEEE Trans. Commun. 69 (2021) 5183–5193, https://doi.org/10.1109/TCOMM.2021.3083597.
[72] S. Kheirinejad, O. Bozorg-Haddad, V.P. Singh, H.A. Loáiciga, The effect of reducing per capita water and energy uses on renewable water resources in the water, food and energy nexus, Sci. Rep. 12 (2022) 7582, https://doi.org/10.1038/s41598-022-11595-w.
[73] H. Pham, A new criterion for model selection, Mathematics 7 (2019) 1215, https://doi.org/10.3390/math7121215.
[74] R. and A.-C.E. American Society of Heating, ANSI/ASHRAE/IES Standard 90.1-2016: Energy Standard for Buildings except Low-Rise Residential Buildings, ASHRAE, 2016.
[75] M.A. Brown, A. Soni, M. V Lapsa, K. Southworth, M. Cox, Low-income energy affordability in an era of US energy abundance, Prog. Energy. 1 (2019) 012002.
[76] S. Sayadi, A. Hayati, M. Salmanzadeh, Optimization of window-to-wall ratio for buildings located in different climates: an IDA-indoor climate and energy simulation study, Energies 14 (2021) 1974, https://doi.org/10.3390/en14071974.
[77] M. Thalfeldt, E. Pikas, J. Kurnitski, H. Voll, Facade design principles for nearly zero energy buildings in a cold climate, Energy Build. 67 (2013) 309–321, https://doi.org/10.1016/j.enbuild.2013.08.027.
[78] K. Brookfield, S. Tilley, Using virtual street audits to understand the walkability of older adults' route choices by gender and age, Int. J. Environ. Res. Publ. Health 13 (2016) 1061, https://doi.org/10.3390/ijerph13111061.
[79] S.J. Mooney, M.D.M. Bader, G.S. Lovasi, K.M. Neckerman, J.O. Teitler, A. G. Rundle, Validity of an ecometric neighborhood physical disorder measure constructed by virtual street audit, Am. J. Epidemiol. 180 (2014) 626–635, https://doi.org/10.1093/aje/kwu180.

[80] Q.C. Nguyen, M. Sajjadi, M. McCullough, M. Pham, T.T. Nguyen, W. Yu, H.-W. Meng, M. Wen, F. Li, K.R. Smith, Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research, J. Epidemiol. Community Health 72 (2018) 260–266.
[81] EIA, 2015 RECS Survey Daa, 2018.
[82] IEA, The Future of Cooling, International Energy Agency, Paris, 2018.
[83] A.M. Fraser, M.V. Chester, D. Eisenman, Strategic locating of refuges for extreme heat events (or heat waves), Urban Clim. 25 (2018) 109–119, https://doi.org/10.1016/j.uclim.2018.04.009.
[84] U.S, Energy Information Administration, Residential Energy Consumption Survey (RECS): 2020, RECE Survey Data, 2020.
[85] J. Kravchenko, A.P. Abernethy, M. Fawzy, H.K. Lyerly, Minimization of heatwave morbidity and mortality, Am. J. Prev. Med. 44 (2013) 274–282, https://doi.org/10.1016/j.amepre.2012.11.015.
[86] J. Zuo, S. Pullen, J. Palmer, H. Bennetts, N. Chileshe, T. Ma, Impacts of heat waves and corresponding measures: a review, J. Clean. Prod. 92 (2015) 1–12, https://doi.org/10.1016/j.jclepro.2014.12.078.