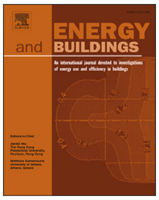
Contents lists available at ScienceDirect

Energy & Buildings

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



Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach

Usman Ali [a,*], Sobia Bano [a], Mohammad Haris Shamsi [d], Divyanshu Sood [a], Cathal Hoare [a], Wangda Zuo [c], Neil Hewitt [b], James O'Donnell [a]

[a] *School of Mechanical and Materials Engineering and UCD Energy Institute, UCD, Dublin, Ireland*
[b] *School of Architecture and The Built Environment, Ulster University, Belfast, UK*
[c] *Pennsylvania State University, University Park, PA, USA*
[d] *Flemish Institute for Technological Research (VITO), Boeretang Mol, Belgium*

ARTICLE INFO

Keywords:
Building energy performance
Data-driven approaches
Urban building energy modeling
Machine learning
Building retrofit

ABSTRACT

Stakeholders such as urban planners and energy policymakers use building energy performance modeling and analysis to develop strategic sustainable energy plans with the aim of reducing energy consumption and emissions from the built environment. However, inconsistent energy data and the lack of scalable building models create a gap between building energy modeling and traditional planning practices. An alternative approach is to conduct a large-scale energy usage survey, which is time-consuming. Similarly, existing studies rely on traditional machine learning or statistical approaches for calculating large-scale energy performance. This paper proposes a solution that employs a data-driven machine learning approach to predict the energy performance of urban residential buildings, using both ensemble-based machine learning and end-use demand segregation methods. The proposed methodology consists of five steps: data collection, archetype development, physics-based parametric modeling, machine learning modeling, and urban building energy performance analysis. The devised methodology is tested on the Irish residential building stock and generates a synthetic building dataset of one million buildings through the parametric modeling of 19 identified vital variables for four residential building archetypes. As a part of the machine learning modeling process, the study implemented an end-use demand segregation method, including heating, lighting, equipment, photovoltaic, and hot water, to predict the energy performance of buildings at an urban scale. Furthermore, the model's performance is enhanced by employing an ensemble-based machine learning approach, achieving 91% accuracy compared to the traditional approach's 76%. Accurate prediction of building energy performance enables stakeholders, including energy policymakers and urban planners, to make informed decisions when planning large-scale retrofit measures.

1. Introduction

The operation of buildings accounted for 30% of global energy consumption and 27% of total energy sector greenhouse gas emissions (GHG) in 2021 [1]. Within this context, 8% comprised direct emissions occurring within buildings, while 19% represented indirect emissions resulting from the production of electricity and heat used in buildings. To address these environmental concerns, the member nations of the European Union (EU) have established a legislative infrastructure to advance sustainable strategic planning initiatives and strengthen energy efficiency within the building sector using the Energy Performance of Buildings Directive (EPBD). The primary objective of this directive is to facilitate the adoption of policies and measures that will enable the achievement of a highly energy-efficient and decarbonized building stock by the years 2030 and 2050, respectively [2].

The rise in annual energy consumption, especially in urban areas, is expected to increase carbon emissions significantly [1]. As a result, there is a growing focus on reducing energy use and emissions from the building sector. Urban planners and policymakers are exploring innovative strategies to make existing buildings more sustainable, in-

* Corresponding author.
E-mail addresses: usman.ali@ucd.ie (U. Ali), sobia.bano@ucdconnect.ie (S. Bano), mohammadharis.shamsi@vito.be (M.H. Shamsi), divyanshu.sood@ucdconnect.ie (D. Sood), cathal.hoare@ucd.ie (C. Hoare), wangda.zuo@psu.edu (W. Zuo), nj.hewitt@ulster.ac.uk (N. Hewitt), james.odonnell@ucd.ie (J. O'Donnell).

https://doi.org/10.1016/j.enbuild.2023.113768
Received 19 September 2023; Received in revised form 1 November 2023; Accepted 17 November 2023

Available online 22 November 2023

Nomenclature

BEM	Building Energy Modeling
BEPS	Building Energy Performance Simulator
BER	Building Energy Rating
CEA	City Energy Analyst
CityBES	City Building Energy Saver
CSO	Central Statistics Office
DEAP	Dwelling Energy Assessment Procedure
DT	Decision Tree
EPBD	European Union Energy Performance of Buildings Directive
EPC	Energy Performance Certificate
GB	Gradient Boosting
HGB	Histogram-Based Gradient Boosting
HVAC	Heating Ventilation, and Air Conditioning
KNN	K-Nearest Neighbor
LGBM	Light Gradient Boosted Machine
LR	Linear Regression
NN	Neural Network
RF	Random Forest
SEAI	Sustainable Energy Authority of Ireland
SVR	Support Vector Regression
UBEM	Urban Building Energy Modeling
UMI	Urban Modeling Interface
XGB	Extreme Gradient Boosting

cluding creating comprehensive sustainable energy plans. Furthermore, long-term renovation strategies are necessary to achieve a higher level of sustainability and reduce carbon emissions from buildings. These plans aim to minimize overall energy consumption and CO_2 emissions by analyzing data on the energy performance of buildings on a large scale. As a result, the EU has implemented the aforementioned EPBD to ensure that member states develop the buildings database comprising Energy Performance Certificates (EPCs). However, even with this mandate, building stock databases typically cover only 30-50% of the total building stock [3].

Moreover, available data are often inadequate for stakeholders such as urban planners, energy policymakers, utility planners, and manufacturers to create effective and sustainable energy conservation measures. Gathering accurate and comprehensive data for urban modeling poses a significant challenge [4]. The limited availability and accessibility of data at the urban scale make it difficult to understand the urban context thoroughly. This poses a hurdle for researchers and practitioners who aim to develop accurate and reliable models that capture the complexities of urban systems. Overcoming this issue requires innovative approaches and collaborations to improve data collection and sharing mechanisms, ensuring a more comprehensive and representative urban modeling and analysis. Similarly, estimating the energy performance of the entire building stock is challenging due to numerous factors that impact energy usage, including the building envelope, the geometry of buildings, the behavior of occupants, heating and cooling systems, and the weather conditions [5,6].

Generally, there are two main approaches to estimating building energy performance: physical and data-driven models [7]. Physical models are based on detailed building physics and are analyzed using simulation tools such as EnergyPlus, ESP-r, and TRNSYS [5]. The simulation of these tools requires extensive building characteristics, including geometric and non-geometric information [6]. On the other hand, the data-driven approach predicts energy usage based on historical data, employing statistical or machine learning algorithms [8]. Unlike the physical modeling approach, this method does not require a deep understanding of the building. This approach has gained significant popularity in the building energy sector because it allows prediction and estimation of energy consumption with limited building information [6]. Similarly, data-driven models can uncover complex relationships between various characteristics of buildings and energy consumption, which can be challenging to identify using traditional methods.

In recent years, researchers implemented various data-driven approaches in building energy demand prediction. These approaches use historical data and employ statistical and machine learning (ML) algorithms to develop data-driven models [6,9–12]. Machine learning algorithms can be broadly classified into supervised and unsupervised learning techniques, with supervised learning further divided into regression and classification algorithms [13]. Supervised learning algorithms commonly used in building energy demand prediction include a nearest neighbor, naive Bayes, rule induction, deep learning, Support Vector Machines (SVM), and neural networks [14,15,13]. On the other hand, unsupervised learning techniques are applied without any corresponding output variable for inputs [14]. Unsupervised learning algorithms commonly implemented in this domain include clustering and association rules of k means [16,11]. However, previous studies employing the data-driven methodology primarily concentrated on forecasting the energy consumption of individual buildings [17]. This limited focus is mainly due to the need for more high-quality and reliable data on a large scale. In addition, these studies have relied on only a few parameters to forecast the potential energy consumption of the building [18].

The novelty of this research lies in the integration of parametric simulations, ensemble-based machine learning approaches, and segregation methods to predict building energy performance at an urban scale using limited resources. Parametric simulation techniques can create synthetic data encompassing a wide range of relevant scenarios for stakeholders. This study implements ensemble-based machine learning algorithms to predict building energy performance on an urban scale by segregating end-use demands such as electricity, hot water, and heating. Furthermore, this research identifies the key building characteristics for each end-use demand prediction. The research additionally analyses the impact of retrofit measures and future stakeholder policies using historical and future weather data.

This paper is structured as follows. Section 2 describes an overview of the existing work done on the prediction of the energy performance of urban buildings. Section 3 outlines the methodology devised, including an explanation of the steps followed in the development of the machine learning model. The results of the Irish case study are presented in Section 4, followed by discussions of possible implications and improvements in the case study in Section 5. Section 6 includes conclusions and potential challenges, and future work.

2. Literature review

Urban building energy modeling can effectively analyze building energy performance and facilitate sustainable energy planning. The most common modeling approaches, such as physics-based or data-driven approaches, differ based on implementation and data requirements, as described in the following sections.

2.1. Physics-based urban building energy modeling

The physics-based urban building energy modeling approach also referred to as the engineering or simulation approach, uses simulation techniques along with data related to building characteristics, construction, weather conditions, and data from heating-cooling systems to compute the consumption of end-use energy [19,20]. The physics-based approach can simulate and estimate building energy usage or production on site, incorporating renewable energy technologies [13]. These models determine the end-use energy consumption of each building by type and rating using measurable data [7].

In the context of cities, the bottom-up archetype method has been widely used to analyze the overall impact of energy efficiency strategies and new technologies at a regional or national scale [5,21]. Each building archetype is modeled in the simulation engine to estimate energy consumption, with these estimates then scaled up to represent the regional or national building stock [22]. These approaches heavily rely on quantitative data obtained from building physics. These methods require various inputs, such as the thermal properties (U values) of the building components (walls, windows, roof, floor, doors), internal and external temperatures, heating system patterns, ventilation rates, appliance quantities, occupancy, schedules, and internal loads [7,6]. In addition, these models require numerous assumptions to establish the behavior of the occupants and a substantial amount of technical data to estimate energy consumption.

One of the most prominent projects, the City Building Energy Saver (CityBES), offers a platform for modeling and analyzing the thermal performance of different retrofit scenarios [23]. CityBES uses the EnergyPlus simulation engine to model buildings and analyze retrofit at the district or city scale [24]. Another project, The CitySim project, involves a decision support tool that assists energy planners and stakeholders in minimizing energy usage and emissions while incorporating various optimization and retrofit analyses [25]. Urban Modeling Interface (UMI) integrates the EnergyPlus simulation engines, Daysim, and a Python module for the operational energy, daylighting, and walkability of urban buildings [26]. MIT's UBEM (Urban Building Energy Model) platform uses the EnergyPlus simulation engine to model approximately 83,541 buildings by integrating official GIS datasets and a custom building archetype library [27]. URBANopt (Urban Renewable Building And Neighborhood Optimization) provides an EnergyPlus and OpenStudio-based simulation software development kit (SDK) to simulate the energy performance of low-energy districts and campus-scale thermal and electrical analyses [28].

One of the significant challenges in modeling at an urban scale is the availability of both building geometric and non-geometric data. Few recent studies have focused on the generation of new building geometric data. UBEM.io, a novel web-based framework, automates the generation of urban-scale building geometries based on widely available inputs such as shapefiles, LiDAR, and tax assessor data [29]. Soroush et al. developed a detailed urban building energy model using the CityGML format for 3D urban geometry and employed spatial joining to incorporate the features required for archetype selection [30]. Ali et al. proposed urban building energy and microclimate modeling by generating 3D city models from sources such as Google Earth, Microsoft Footprints, and OpenStreetMap [31]. Irene et al. developed a modeling framework to assess the potential of creating energy communities by combining UBEM capabilities with the rooftops' potential for solar generation [32].

With increased data availability and more sophisticated modeling techniques, it has become crucial to devise a generalized UBEM framework and improve the existing work to facilitate the modeling and analysis of different use cases. Previous studies provide a limited view of the different building energy aspects in an urban setting. This stems mainly from the fact that simulating each building individually, along with their interdependencies, requires significant time and resources [33]. Furthermore, these methods usually deploy a physics-based simulation engine, which can be computationally demanding and time-consuming due to the intricate nature of urban systems".

Data-driven urban building energy modeling can address the aforementioned challenges by estimating building energy consumption using basic knowledge of the buildings' features. However, this approach still has research gaps, as discussed in the next section.

2.2. Data-driven urban building energy modeling

In urban energy modeling, a data-driven approach can predict and assess buildings' energy usage by considering various factors related to the characteristics of the buildings [7,19]. This approach is based on the analysis of existing data sources that include building stock datasets, billing data (such as electricity and gas consumption), survey data, and socioeconomic variables [7]. Data-driven urban energy modeling is conducted mainly using machine learning and statistical approaches. Recent studies on urban energy have increasingly focused on using machine learning algorithms over traditional statistical techniques [7].

Rahman et al. used deep recurrent neural networks to predict medium- to long-term electricity use in commercial and residential buildings [34]. Meanwhile, Kontokosta and Tull devised statistical models to determine the energy consumption of electricity and natural gas in more than a million buildings in New York City [35]. Feifeng et al. proposed a semi-supervised learning method for predicting energy use intensity (EUI) using 34,456 unlabeled samples [36]. Zhang et al. proposed a data-driven framework for the prediction of energy usage and greenhouse gas emissions, which considered various factors such as building characteristics, geometry and urban morphology [37]. Similarly, Seo et al. developed a data-driven model to predict the energy demand for heating of 10,000 low-income households in South Korea [38]. Razak et al. developed a machine learning model that forecasts annual average energy use based on building design features in the initial development stages [18]. Ngo et al. used ensemble machine learning models to forecast building energy consumption over 24 hours [39]. Lastly, Wurm et al. developed a workflow for modeling the heat demand of building stock on an urban scale, using deep learning algorithms [40].

Although a significant amount of research has been conducted on predicting energy consumption in individual buildings using their specific characteristics, more studies have yet to explore using data-driven models for predicting energy consumption on a larger scale. The main challenge lies in the lack of high-quality data in sufficient quantities to train prediction models effectively. This underscores the need for a robust building energy modeling approach capable of accurately predicting the energy performance of entire building stocks, even when faced with limited resources for complex decision-making analysis. Furthermore, previous research on predicting building energy consumption has been limited by considering only a small set of parameters ([18]). Fewer recent studies have started incorporating crucial factors such as U-values, HVAC systems, and renewable energy systems into their machine-learning algorithms to estimate better energy performance in buildings ([37]). However, only a few studies have specifically investigated the impact of parameters such as U values, HVAC system types, and the presence of renewable energy systems on the estimation of the energy performance of buildings using machine learning algorithms ([18,39–41]).

Predicting the energy performance of buildings at an urban scale poses a significant challenge for urban planners and policymakers. The accurate prediction of energy consumption and the identification of opportunities for enhancing energy efficiency are crucial for fostering sustainable development in cities. There is significant potential to expand current research and establish a comprehensive methodology for data-driven building energy modeling on an urban level.

However, one major issue that arises in an urban context is the availability of data. Obtaining comprehensive and reliable data at an urban scale can be challenging, as it requires collecting and integrating information from multiple sources [4]. Addressing this issue is essential to enable effective energy planning and modeling techniques, empowering stakeholders to make informed decisions and drive positive change in urban energy management.

These findings highlight the importance of adopting a holistic approach to building energy modeling, considering all relevant factors, to accurately predict building energy performance and align with the

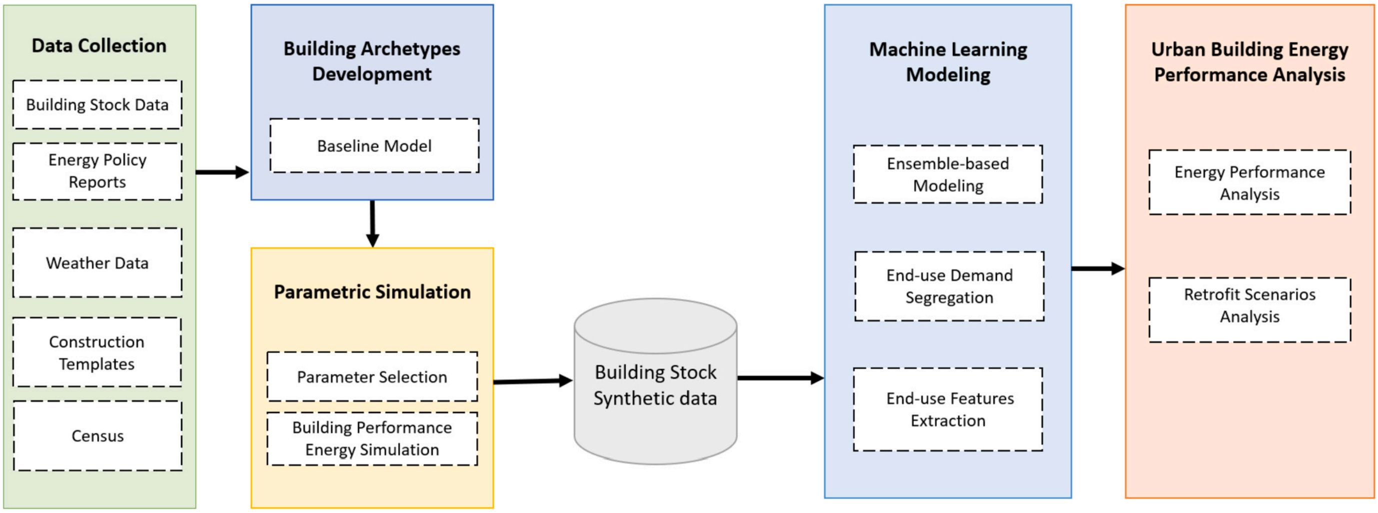

Data Collection
Building Stock Data
Energy Policy Reports
Weather Data
Construction Templates
Census
Building Archetypes Development
Baseline Model
Parametric Simulation
Parameter Selection
Building Performance Energy Simulation
Building Stock Synthetic data
Machine Learning Modeling
Ensemble-based Modeling
End-use Demand Segregation
End-use Features Extraction
Urban Building Energy Performance Analysis
Energy Performance Analysis
Retrofit Scenarios Analysis

Fig. 1. Overarching methodology for urban building energy performance prediction using machine learning.

objectives of various stakeholders. Therefore, this research proposes a methodology that combines and harnesses the strengths of physics-based and data-driven approaches to accurately predict the energy performance of buildings on an urban scale. In the physics-based approach, parametric simulation methods are employed to generate synthetic data that encompass all possible scenarios relevant to stakeholders. Similarly, ensemble machine learning and end-use demand segregation methods are used in the data-driven approach instead of relying on a single model to achieve accurate predictions of building energy performance on an urban scale.

3. Methodology

This study proposes a novel methodology that uses supervised machine learning algorithms to predict building energy performance on a large scale. This research aims to identify the most effective model using physics and data-driven approaches. The prediction methodology for the energy performance of urban buildings involves five steps (Fig. 1).

1. The initial step involves collecting data from various sources such as building stock, census, weather, and geographical data.
2. The next step involves developing building archetypes using existing building stock data to identify representative baseline models.
3. The subsequent step focuses on parametric simulation to develop appropriate synthetic data.
4. The step of developing machine learning models predicts building energy performance on a large scale using an ensemble or segregation method.
5. Finally, the urban building energy performance analysis step analyzes the modeling process results for planning and decision-making purposes.

3.1. Data collection

The data collection process involves gathering various inputs for urban building energy performance prediction using machine learning, including building stock data, weather information, census data, reports on energy policies, and construction data [5].

The building stock data are necessary for conducting physics-based simulations that encompass buildings' geometry and non-geometry data. This includes data such as building envelope specifications, shapes, number of floors, type of building, geometry, geographical position, and window opening ratios ([42]). Typically, the geometric data required for building energy modeling is gathered from building stock and energy performance certificate databases and existing construction databases such as TABULA, EPISCOPE, and building typology databases ([43]).

Along with geometric data, non-geometric data are also required for modelings, such as user occupancy patterns, equipment loads, HVAC systems, and usage patterns also need to be modeled. One of the significant challenges in this regard is the availability of non-geometric building information on a large scale. Non-geometric building data can be obtained through the building archetypes approach, using available national census databases, statistical surveys, and energy performance certificate data.

Weather data sets are essential to accurately model energy use in building thermal simulations ([44]). The most commonly used climate data sets, such as the typical meteorological year data (TMY), have been available for a long time and describe the local climate ([45]). Another helpful resource are EnergyPlus Weather format (EPW) files, which can be accessed online for more than 3,034 locations. These files are arranged by region and country of the World Meteorological Organization. Furthermore, this study incorporates future weather files to assess the impact of weather conditions on retrofit measures under various climate scenarios, aiming to achieve the energy policy targets set by policymakers, such as those for 2030 or 2050. The sources of these future weather files can vary, including resources like Meteonorm, WeatherShift, and CCWorldWeatherGen [46].

Similarly, the modeling process relies on additional sources such as census data, reports on energy policies, and construction data. These sources offer valuable insights into demographic patterns, energy consumption trends, and infrastructure development, facilitating a more comprehensive analysis and meeting the requirements of urban systems.

3.2. Building archetypes development

Several buildings on an urban scale often share similar characteristics and can be classified into building archetypes. In the context of urban building energy simulation, a building archetype, referred to as a reference building, is a representative model that captures the typical characteristics and performance of a specific category or group of buildings within a large building stock. The parametric simulation framework uses each building archetype as a baseline model. These data can be sourced from established national building stock databases, such as the TABULA or EPC databases [43]. Building archetypes or reference buildings serve as standardized models that simplify the simulation process by providing a baseline or template for analysis. They are typically developed based on existing data collection, statistical analy-

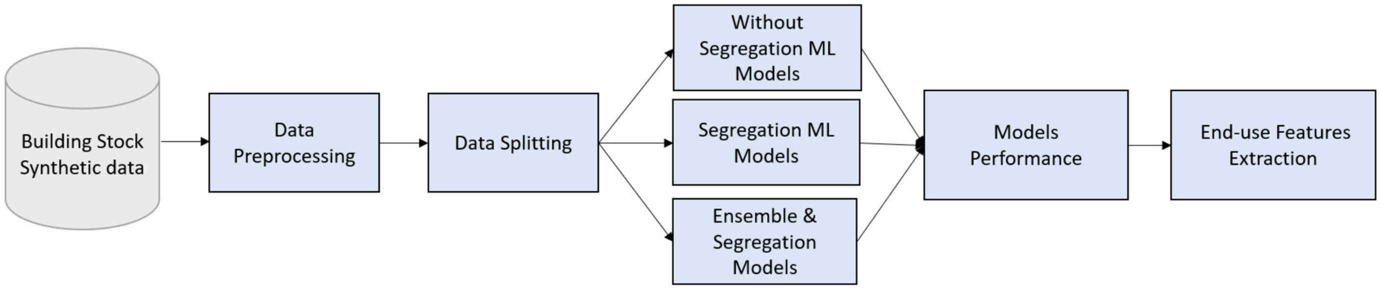

Building Stock Synthetic data
Data Preprocessing
Data Splitting
Without Segregation ML Models
Segregation ML Models
Ensemble & Segregation Models
Models Performance
End-use Features Extraction

Fig. 2. Process of machine learning modeling to predict Energy Use Intensity (EUI) using machine learning models.

sis, and empirical studies of buildings within the target building stock. Moreover, simulating any building archetype requires geometric and non-geometric data for each baseline model. These building archetypes are the starting point for parametric modeling of different buildings to develop a synthetic stock.

3.3. *Parametric simulation*

Parametric simulation provides an optimal solution, mainly when only sparse data sets are available for energy modeling. To execute complex parametric simulations involving multiple parameters, a parametric tool is used to perform numerous simulations using a Building Energy Performance Simulator (BEPS) model ([47]). This study uses jEPlus as a parametric tool for energy simulations. Furthermore, jEPlus uses EnergyPlus for simulation and incorporates DesignBuilder construction templates to integrate diverse parameter values. Parametric simulation using EnergyPlus presents a robust approach to assess the energy performance of buildings and investigate various design alternatives. In the parametric simulation, EnergyPlus facilitates a systematic exploration of the design parameters, providing insights into their impact on energy consumption, comfort, and other performance metrics.

The selection of parametric features plays a crucial role in developing parametric simulation-based models and generating synthetic datasets. The accuracy of the building energy model is highly dependent on the careful selection of each parameter in this process. These parameter values, which encompass the necessary variations for synthetic data generation, can be obtained from literature surveys that are specific to the relevant climate environments ([48,3]).

In the parametric simulation process, various essential parameters are commonly used that include construction characteristics such as walls, windows, floors, roofs, internal gains, occupancy density, and heating or cooling systems. They all contribute to the overall energy performance assessment and are integral to the parametric simulation. By considering these parameters and their variations, parametric simulation enables the exploration of different design alternatives and their impact on energy consumption, comfort levels, and other performance metrics. It allows for a comprehensive evaluation of the energy efficiency of the building and helps to make decisions about design optimizations. Therefore, selecting the appropriate parameters and their values, based on literature surveys and specific climate environments, is crucial to create accurate and representative synthetic datasets and ensuring the reliability of parametric simulation-based models.

However, dealing with the complexity of many parameters makes it nearly impossible to generate simulated data for all possible combinations. Sampling methods such as Simple Random Sampling (SRS) and Latin Hypercube Sampling (LHS) are used to generate synthetic data to address this challenge ([49,50]). Simple Random Sampling (SRS) is a straightforward method in which each sample is randomly and independently selected from the population. On the other hand, Latin Hypercube Sampling (LHS) is a more advanced sampling method that aims to achieve a more uniform distribution of samples across the entire range of the data. LHS ensures that each parameter value combination is balanced, allowing for a more comprehensive design space exploration. These methods allow for generating representative synthetic datasets encompassing a range of parameter combinations, facilitating a more comprehensive analysis of design alternatives and optimizing energy modeling outcomes.

3.4. *Machine learning modeling*

This process involves formulating machine learning models to estimate the building energy performance (Fig. 2). Synthetic building stock data, generated from the parametric simulation step, is intended to serve as input for the development of machine learning models.

3.4.1. *Data preprocessing*

The process begins with data preprocessing, during which inconsistencies within the dataset are identified and eliminated before the data are used for further analysis and model development.

3.4.2. *Data splitting*

The pre-processed data is divided into two subsets to ensure optimal training of the model: a training dataset used for training the model and a test dataset for evaluating the performance of the trained model. Two standard techniques for data splitting are random data splitting and cross-validation.

Random data splitting is a straightforward method in which data is randomly divided into training and testing datasets, typically in an 80-20% split ratio. However, this method may cause problems with uneven data distribution, and an incorrect selection of training and testing datasets can also adversely affect the machine learning model's performance [51]. On the other hand, cross-validation is a more sophisticated method that is often used to strike a balance between minimal bias and variance in the trained model. This study adopts the k-fold cross-validation algorithm for data splitting to prevent overfitting or underfitting the model.

3.4.3. *Non-segregation models development*

This paper implements and compares three different machine learning model approaches to predict building energy performance, namely: the single model approach, end-use demand segregation method, and ensemble-based segregation method. In the single model approach, also referred to as the "non-segregation" method, this study conducts a comparative analysis of various machine learning algorithms, assessing their predictive accuracy, efficiency, and suitability for building energy performance modeling. Over recent years, machine learning models have garnered considerable attention in data-driven modeling. Among the most frequently used models are Linear Regression (LR), Neural Network (NN), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting (GB) and Support Vector Regression (SVR) [7]. Some of the popular implementations of gradient boosting include XGBoost (Extreme Gradient Boosting), Histogram-Based Gradient Boosting (HGB), and LGBM (Light Gradient Boosted Machine). These algorithms have demonstrated exceptional performance in energy forecasting and prediction, particularly in the context of energy modeling, due to their extensive use and success in previous studies ([17,11]). By

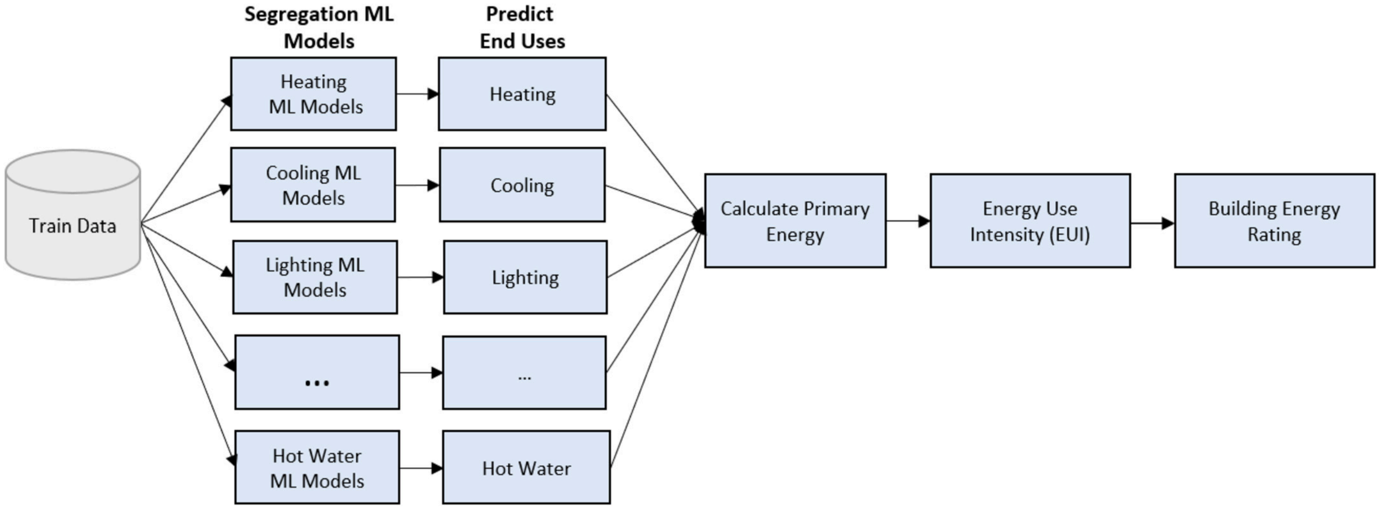


Fig. 3. Methodology for end-use demand segregation modeling to predict Energy Use Intensity (EUI) using machine learning.

assessing the effectiveness of these models, this study aims to discern the most efficient approach to predict building energy performance using machine learning techniques.

3.4.4. End-use demand segregation models development

End-use demand segregation methods use different machine learning models to predict each end-use demand. This strategy diverges from the traditional approach of employing a single machine-learning model. This modification aims to achieve superior predictive performance (Fig. 3). The workflow includes developing distinct regression machine learning models for each end-use demand, such as heating, cooling, lighting, and hot water. The predictions of these end-use demands are aggregated to calculate the final energy performance of the building, measured in terms of Energy Use Intensity (EUI). The prediction for each end-use demand is multiplied by its corresponding primary energy factor. The resulting values for heating, cooling, equipment, lighting, and hot water are then aggregated and photovoltaic energy generation is deducted from them to calculate the total energy consumption of the building. This cumulative total is then divided by the building area to calculate the Energy Use Intensity (EUI), a measure of the energy performance of the building as defined in Equation (1). Finally, the EUI is classified into an Energy Performance Certificate (EPC) label or rating,

$$EUI = \frac{(E_{\text{heating}} \times PEF_{\text{heating}}) + (E_{\text{cooling}} \times PEF_{\text{cooling}})}{A_{\text{total}}} + \frac{(E_{\text{lighting}} \times PEF_{\text{lighting}}) + (E_{\text{equipment}} \times PEF_{\text{equipment}})}{A_{\text{total}}} + \frac{(E_{\text{hotwater}} \times PEF_{\text{hotwater}}) - (E_{\text{PV}} \times PEF_{\text{PV}})}{A_{\text{total}}} \quad (1)$$

where $E_{heating}$, $E_{cooling}$, $E_{lighting}$, $E_{equipment}$, $E_{hotwater}$, and E_{PV} represent the energy consumption (or generation for E_{PV}) for each respective category in kilowatt hours per year (kWh/year).$PEF_{heating}$, $PEF_{cooling}$, $PEF_{lighting}$, $PEF_{equipment}$, $PEF_{hotwater}$, and PEF_{PV} are the primary energy factors (PEFs) for each respective category. A_{total} represents the total floor area of the building in square meters (m^2).

3.4.5. Ensemble and segregation models development

The workflow further implements ensemble machine learning methods to test multiple learning algorithms and obtain better predictive performance. Ensemble techniques are commonly used in machine learning to enhance model accuracy by mitigating overfitting and increasing generalizability. By leveraging the complementary strengths of multiple models, ensemble learning provides more stable predictions and improves accuracy compared to the conventional approach of using a single model. There are two main ensemble learning techniques that differ mainly by kind of model, data sampling, and decision function. Therefore, ensemble learning techniques can be classified as stacking and voting techniques.

The stacking method, also known as stacking generalization, was introduced by Wolpert [52]. The goal is to reduce the generalization error of different machine learning models. The final Meta-Model comprises the predictions of an "n" number of machine learning-based models through the k-fold cross-validation technique. On the other hand, the voting ensemble method is one of the most intuitive and easy to understand. The voting ensemble method comprises a number "n" of machine learning models, and the final prediction is the one with "the most votes" or the highest weighted and averaged probability. Generally, ensemble learning techniques use multiple best-prediction performance machine learning models. The study implements a stacking-based ensemble method to predict each end-use demand, enhancing model accuracy and predicting building energy performance. This method combines predictions from multiple models by training another model to consolidate its output, often resulting in more accurate and robust predictions compared to the voting ensemble method (Fig. 4).

3.4.6. Models performance

To evaluate the effectiveness of machine learning models, commonly used performance indices such as R-Squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are employed ([7,11]). A model with the lowest RMSE and MAE values and a R^2 value nearest to 1 is deemed superior among all models. Finally, in order to assess the model's accuracy, the predicted value of EUI (expressed in kWh/(m^2*year)) is transformed into an Energy Performance Certificate (EPC) label or rating. Furthermore, precision and recall are crucial metrics used for a detailed analysis of each class. Precision assesses the accuracy of positive predictions made by the model, whereas recall quantifies the model's capability to detect all positive instances within the dataset [3].

3.4.7. End-use features extraction

The final step of this process is to find the importance of features for each end-use demand using the developed machine learning model. Feature importance refers to the determination of the relevance or contribution of individual features in a machine learning model to make accurate predictions. It helps in understanding which features have the most significant impact on the model's predictions.

One popular method for calculating feature importance is SHAP (SHapley Additive exPlanations). SHAP values provide a unified mea-

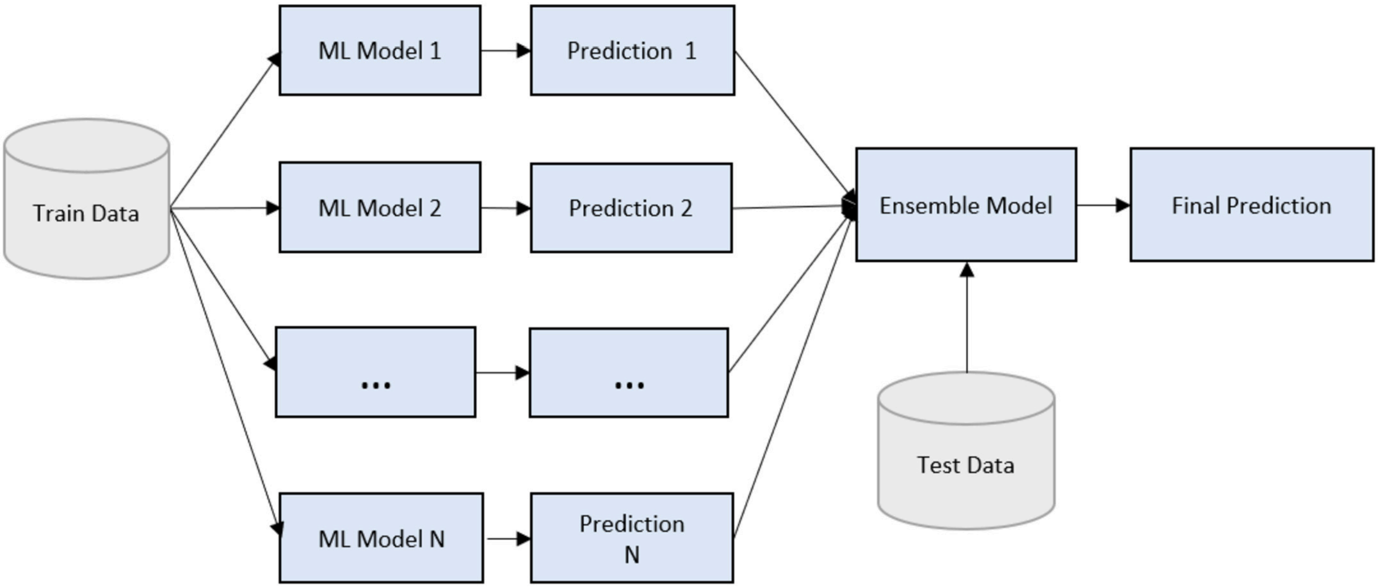


Fig. 4. Methodology for ensemble machine learning modeling approach for enhanced predictive performance in machine learning models.

sure of feature importance by considering the contribution of each feature value to the prediction for a specific instance while also accounting for interactions between features. By using SHAP values, we can gain insight into which features impact the model's predictions the most. This information can be valuable for understanding the underlying relationships in the data and identifying the key drivers or factors that influence the target variable.

3.5. Urban building energy performance analysis

In the final phase of the methodology, the developed machine learning model predicts the energy performance of the entire building stock. The availability of comprehensive building stock data can help stakeholders analyze the building stock at an urban scale and successfully implement sustainable energy policies. Furthermore, the developed model can be applied to practical application scenarios, such as implementing and evaluating proposed retrofit measures as part of national-level policy decisions. These measures, often proposed at the national level, aim to improve the energy performance of existing buildings through modifications and improvements. For example, this could include installing heat pumps or integrating renewable energy systems like solar panels. The proposed models can evaluate their impact before implementation and identify potential energy savings. This predictive capability reduces the risk of implementing ineffective or inefficient measures, ensuring that resources are used optimally. It also helps fine-tune such measures to fit better the specific needs and constraints of the building stock.

In general, the developed model offers a holistic approach to urban-scale energy management and policy implementation, creating a more sustainable built environment. Using modeling outcomes, stakeholders can navigate the complexities of urban building stock analysis and energy policy implementation, even without extensive knowledge of building dynamics. This empowers policymakers and stakeholders alike to make informed decisions when retrofitting existing building stock to improve energy efficiency and mitigate environmental impact.

4. Case study

The primary objective of this case study is to test the proposed methodology by calculating the energy performance of Ireland's residential building stock. This methodology seamlessly integrates a data-driven approach with parametric simulation modeling to predict the energy performance of buildings on an urban scale. This case study follows the same structure as the proposed methodology discussed in the previous section, with subsequent subsections following the same order.

4.1. Data collection

Collecting urban-scale building stock data is challenging as individual building information is often unavailable [4]. The data collection process involves acquiring raw building data from various sources to implement the proposed methodology, including building stock datasets, building census datasets, weather data, and data from energy policymakers' reports. See Table 1.

In Ireland, building stock data are available as Energy Performance Certificates (EPCs) maintained by the Sustainable Energy Authority of Ireland (SEAI). The EPC (also called the Building Energy Rating (BER) certificate) dataset of the Irish residential stock represents the measured building stock and comprises more than 200 building characteristics. These features include building fabric, heating systems, estimated end-use, CO_2 emissions, and estimated delivered and primary energy consumption. Each entry in the Irish EPC dataset contains an energy rating for the respective building, ranked its energy performance on a graded scale from A1 to G based on the estimated energy consumption per square meter per year [53]. In 2023, the Irish EPC dataset contained approximately 1,126,817 residential buildings, with a significant proportion of building ratings within the range of C1 to D2 (Fig. 5). The dataset's most common types of buildings are semi-detached and detached houses.

The Irish census, conducted every four years by the Central Statistics Office (CSO), collects various data points on the building where the respondent resides. Therefore, the census provides the number of buildings in each geographic area [56]. According to the CSO 2022 dataset, Ireland has approximately 1,841,152 residential buildings. Similarly, the GeoDirectory database provides statistical and geographical information on Ireland's entire building stock [54]. The Q4 2022 GeoDirectory report, published by An Post (Irish Postal Service) and Ordnance Survey Ireland, comprises geocoded addresses of 2,100,905 residential buildings in Ireland. Detached dwellings remained the most prevalent type of residence (30.7% of the national total), followed by terraced dwellings (28.2%) and semi-detached dwellings (24.7%). This study focuses on Dublin City in Ireland and the Dublin EPC dataset, which includes 339,494 of the 624,758 residential buildings, representing the highest proportion of the entire Irish building stock. This suggests that

Table 1
Building data requirements and associated data sources for Irish case study.

Data Type	Case Study Data Source	Publisher
Building Stock	Irish EPC (BER) Database [53]	SEAI
Geographic data	GeoDirectory [54,55]	An Post/ Ordnance Survey Ireland
Census	Irish Cenus database [56]	Central Statistics Office
Weather	Dublin EPW File [45]	EnergyPlus and Meteonorm
Energy policymakers' Reports	Irish Climate Action Plan [57]	Government of Ireland

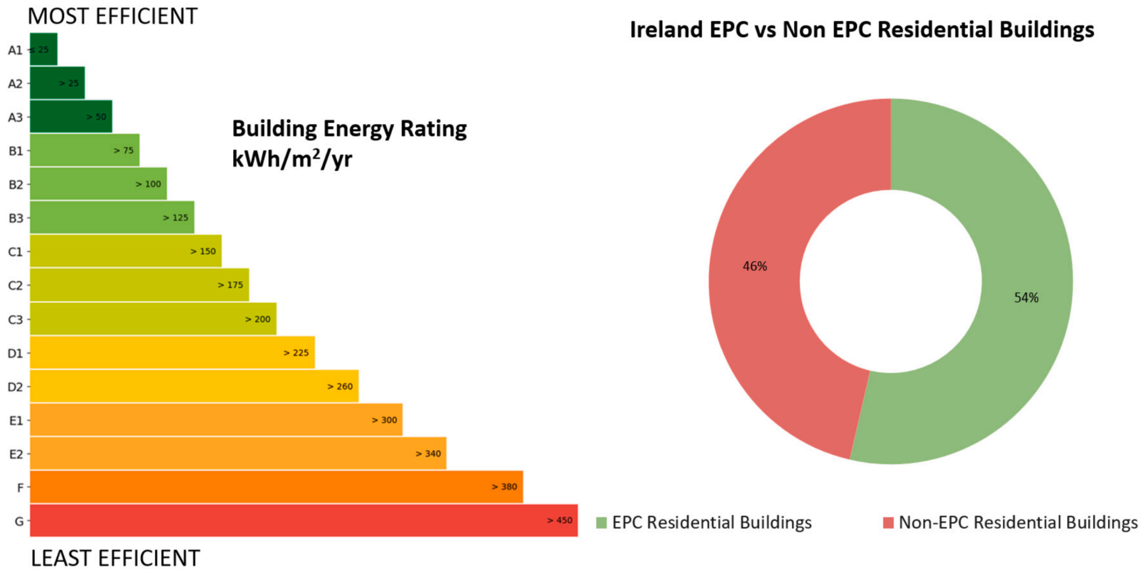
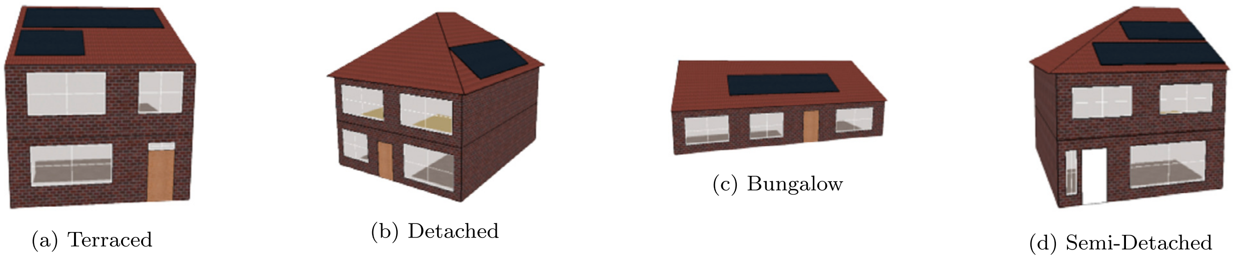


Fig. 5. Irish EPC building energy rating chart used to determine building energy performance, percentage of total EPC vs. Non-EPC residential buildings.


(a) Terraced

(b) Detached

(c) Bungalow

(d) Semi-Detached

Fig. 6. 3D geometry of Irish residential building archetypes for energy parametric simulation [44,48].

EPC data are available for only approximately 54% of the residential building stock of Dublin City ([53]). This study employs machine learning algorithms to predict the energy rating of the remaining 46% stock using limited variables (Fig. 5). Furthermore, the weather data for Dublin are obtained from the default EnergyPlus dataset, which includes historical data and also incorporates future weather files for 2030 by Meteonorm. This allows us to assess the impact of weather conditions on retrofit measures in various climate scenarios.

Similarly, energy policy reports are necessary to explore future scenarios. Irish national reports, such as the Climate Action Plan 2023, are used to test scenarios in this case study. This provides valuable insight into future plans and strategies for Irish residential buildings. These reports outline the goals, roadmaps, and goals set by policymakers to address climate change, reduce greenhouse gas emissions, and improve energy efficiency in the residential sector [57].

4.2. Building archetypes development

The parametric simulation framework uses each building archetype as a baseline model. In this case study, four building types are considered as archetypes of the Irish residential building stock [44]. These types are selected to represent the primary variations of building types based on data from the CSO, Irish EPC, and GeoDirectory datasets. These building archetypes serve as the starting point for the parametric modeling of different buildings, helping to develop a synthetic stock representation. These four different types of residential buildings also exist in the GeoDirectory database, namely terraced houses, detached houses, semi-detached houses, and bungalows (Fig. 6).

Building archetypes require both geometric and non-geometric data to model each baseline model. The initial step involves identifying the non-geometric and geometric parameters associated with the existing building stock of Dublin. This information is essential for performing a parametric simulation using the archetypes. Geometric information collected from various types of Irish buildings is based on existing studies and Irish building regulations guidelines. However, non-geometric parameters are determined using current building energy performance databases and literature surveys. For example, the Irish EPC provides values for essential building physics parameters, such as U-values for walls, roofs, floors, and windows, along with their respective ranges. Other relevant non-geometric parameters that impact the energy performance of the Irish building stock have been identified based on previous research [44,48]. The geometric and non-geometric parameters of base-

Table 2
Geometric and non-geometric parameters of baseline archetypes used in the Irish case study.

Geometric Parameters (Default Model Values)					
Parameters	Unit	Terraced	Detached	Semi-detached	Bungalow
Total Floor Area	m^2	91.66	130.81	107.69	85.91
Net Conditioned Area	m^2	91.66	130.81	107.69	85.91
Gross Roof Area	m^2	65.66	115.68	81.76	130.43
Window to Wall Ratio on NWSE facades	%	0.4/0/0.4/0	0/0.5./0/0.5	0.4/0/0.4/0	0.4/0/0.4/0
Number of Stories (Height 2.7 meters)	Numeric	2	2	2	1
Number of Zone	Numeric	10	13	10	8
Orientation	degree	0	90	0	0
Non-GeometricParameters (Default Model Values)					
Wall U-value	W/m^2 K	0.5	0.5	0.5	0.5
Window U-value	W/m^2 K	3	3	3	3
Floor U-value	W/m^2 K	0.5	0.58	0.5	0.58
Roof U-value	W/m^2 K	0.33	0.33	0.33	0.33
Door U-value	W/m^2 K	2.041	2.041	2.041	2.041
Lighting Density	W/m^2	2.92	2.95	2.92	3.025
Occupancy	Person	3	4	3	4
Equipment Density	W/m^2	1.47	1.61	1.47	1.56
Heating setpoint	˚C	21	21	21	21
Heating setback	˚C	12	12	12	12
HVAC Efficiency/ COP	Numeric	0.8	0.8	0.8	0.8
DHW	l/m^2/day	1.5	1.5	1.5	1.5
ACH	Numeric	0.94	0.87	0.94	0.74
Renewables	W	2400	2400	2400	2400

Table 3
Parameters needed for parametric simulation of archetypes.

No	Parameters	Unit	Minimum	Maximum	Source
P1	Building type	Categorical	Semi Detached, Detached, House, Terrace, Bungalow		[53]
P2	Location	Categorical	Dublin		[56]
P3	Weather	Categorical	Historical, 2030		EPW
P4	Wall U-value	W/m^2 K	0.09	2.4	[48,58]
P5	Window U-value	W/m^2 K	0.73	5.7	[48,58]
P6	Floor U-value	W/m^2 K	0.15	1.23	[48,58]
P7	Roof U-value	W/m^2 K	0.07	2.3	[48,58]
P8	Door U-value	W/m^2 K	0.81	5.9	[48,58]
P9	Orientation	degree	0	315	[48,58]
P10	Lighting density	W/m^2	1	9	[48,58]
P11	Occupancy	Person	1	6	[56]
P12	Equipment density	W/m^2	1	21	[48,58]
P13	Heating setpoint	˚C	18	23	[48,58]
P14	Heating setback	˚C	10	14	[48,58]
P15	HVAC efficiency or COP	0.45 to 4	0.3	4.5	[53]
P16	Domestic hot water	l/m^2/day	0.5	3.5	[48,58]
P17	Air changes per hour	Numeric	0.35	3	[59,53]
P18	Window-to-wall ratio	%	30	70	[48,58]
P19	Renewables	W	Yes/No		[53]

line archetypes with default values used for the Irish case study are shown in Table 2 [44,48,62,63].

4.3. Parametric simulation

The selection of parametric features is pivotal in developing physics-based models based on parametric simulation and generating synthetic datasets after the archetype development process. The accuracy of the building energy model relies on the careful selection of each input and output parameter in this process. These parameter values embody the necessary variations for synthetic data generation. In this study, 19 input parameters are used to simulate Irish residential building archetypes. The selection of these parameters is based on existing studies on residential buildings [48,3]. However, these previous studies do not include certain advanced features. Therefore, several additional parameters, including HVAC systems, are incorporated to conduct a complete analysis of HVAC systems, primary heating factors, and renewable parameters (Table 3). Furthermore, this study employed a building feature reduction approach by integrating Design-Builder construction templates and reducing the number of dependent features. For instance, building elements require material features such as thickness, conductivity, density, and specific heat. In this study, existing templates were used, and U-values were used to represent these features. This approach ultimately results in a reduction of the required parameters as inputs to the UBEM and further reduces the model computing time by eliminating dependent parameters.

One of the primary output parameters in this study is the Energy Use Intensity (EUI), also referred to as the final primary energy use per building's total floor area per year, measured in kW h/(m^2*year). Irish EPC data provide information on building energy performance or certificate ratings in terms of EUI (kW h/(m^2*year)), which is further interpreted on an A1 to G rating scale. An A1-rated building demonstrates the highest level of energy efficiency, typically associated with the lowest energy consumption and CO_2 emissions. On the other hand, a building with a G rating represents the least energy-efficient rating (Fig. 5). Furthermore, this study focuses on the end-use demand segregation method to calculate the Energy Use Intensity. Therefore, each

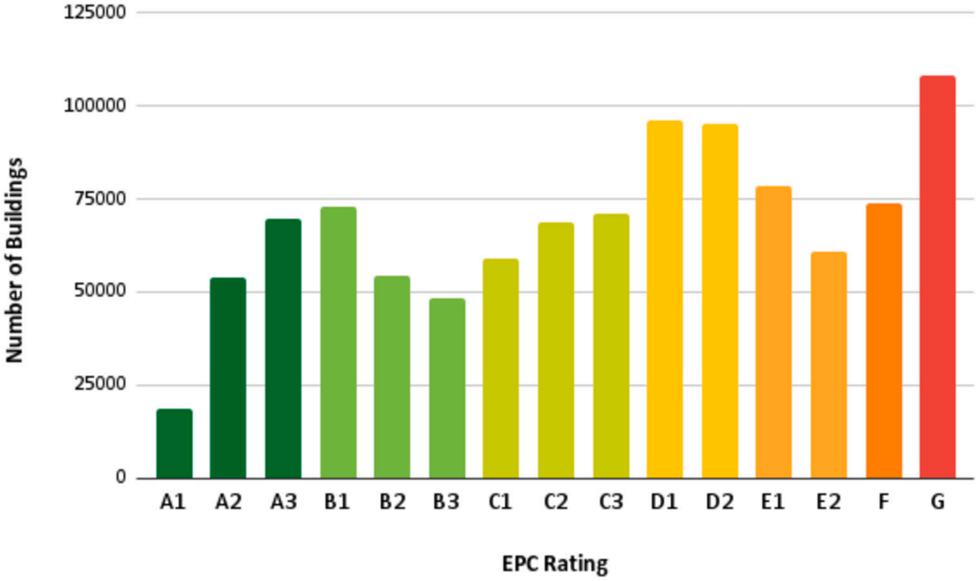


Fig. 7. Distribution of 1 million residential buildings synthetic data in terms of the Irish building energy rating labels.

Table 4
Comparative analysis of machine learning models to predict end-use demand in kW h/yr using RMSE metrics.

Models	Heating	Interior Lighting	Interior Equipment	Photovoltaic Power	Water Systems
XGB	683.17	0	0	0.02	0
LGBM	801.69	0	0	0	0
HGB	1256.58	0.02	0.06	0.06	0.21
GB	2809.86	67.72	193.83	16	13.56
RF	1613.23	0	0	0	0
NN	3400.93	1.01	18.94	6.85	16.78
DT	2430.7	0	0	0	0
LR	5162.23	181.24	546.94	172.78	6440.26
KNN	5106.97	175.26	483.64	310.06	5629.45
SVM	7330.98	192.2	575.37	175.56	7976.76

end-use demand, including heating, lighting, equipment, photovoltaic, and hot water, is considered an output parameter in the parameter simulation process.

This study employs jEPlus as a parametric tool for physics-based parametric simulation. A jEPlus uses the capabilities of EnergyPlus for thermal simulation and integrates DesignBuilder construction templates to incorporate diverse parameter values. A sample of 1 million buildings is generated using the Latin hypercube sampling (LHS) method to construct a reliable machine learning model. This sampling process ensures that the resulting distribution covers all energy rating data for Irish buildings (Fig. 7).

4.4. Machine learning modeling

This process involves formulating an urban-scale building energy performance machine learning model. The process begins with generated synthetic building stock data from the previous step, which are preprocessed to remove outliers and improve the data set's quality before implementing machine learning models. Subsequently, the data is divided into two subsets to create training and testing datasets. This study uses a 10-fold cross-validation method during data division to mitigate the risk of overfitting, rather than using a random data selection for training and testing.

Ten different machine learning algorithms are analyzed to assess their abilities to predict EUI building energy performance based on a given dataset. These regression algorithms have shown exceptional performance in energy forecasting and prediction, particularly within the context of energy modeling ([17,11,7]). The algorithms include XGBoost (XGB), LightGBM (LGBM), Gradient Boosting (GB), Histogram-based Gradient Boosting (HGB), Random Forest (RF), Neural Network (NN), Decision Tree (DT), Linear Regression (LR), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The performance of each developed model is evaluated using metrics such as R-Squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). A model is considered superior if it achieves values closer to zero for RMSE and MAE and values close to zero for R^2. The target feature is EUI, which is used to predict building energy performance using regression models. Furthermore, the final predicted EUI is also converted into an energy rating based on the Irish EPC rating (Fig. 5). Finally, the model's performance is further tested using an accuracy estimation of the energy rating, with the model producing the highest accuracy being considered the best learning model.

This study conducts a comparative analysis of three different machine learning models proposed in this research to evaluate which one is best suited for predicting building energy performance. These approaches include the single-model approach (non-segregation method), the end-use demand segregation method, and the ensemble-based segregation method. In the non-segregation method, EUI predicted using all ten machine learning models. Similarly, the workflow then develops learning models using the segregation method for each end-use demand, such as heating, interior lighting, photovoltaic power and water systems in the interior equipment. The process implemented and tested ten machine learning models for each end-use demand (Table 4). The results show that the XGB model showed the best performance in predicting the demand for heating with an RMSE of 683.17. For interior lighting, interior equipment, photovoltaic power and water systems, the XGB,

Table 5
List of important features with rank that affect end-use demand machine learning models using SHAP method.

Rank	Heating	Lighting	Equipment	Photovoltaic	Water Systems
1	Air changes per hour	Lighting density	Equipment density	Renewables	Building type
2	Heating setpoint	Building type	Building type	Orientation	Domestic hot water
3	Wall U-value			Weather	
4	Building type				
5	Occupancy				
6	Window U-value				
7	Equipment density				
8	Weather				
9	Roof U-value				
10	Lighting density				
11	Heating setback				
12	Floor U-value				

LGBM, RF and DT models reported an RMSE of 0, indicating excellent performance.

In addition, models such as LR, KNN, and SVM exhibited relatively higher root mean square errors (RMSE) in all categories, indicating less accurate predictions. The results demonstrate that the RMSE for most end-use demands is nearly 0. This can be attributed to the fact that end-use demands calculated in EnergyPlus are derived using static calculations, meaning that values are determined based on fixed parameters and equations without accounting for variability or randomness. Therefore, machine learning models can easily learn and map these fixed relationships between input features and end-use demands, resulting in a near-perfect fit to the data. Furthermore, the SHAP method is employed to gain further insight into the main features that affect the model output (Table 5). The findings reveal significant factors that affect energy consumption in buildings. The rate of air changes per hour emerged as the most influential feature, highlighting the importance of ventilation in determining heating demand. The heating setpoint and wall U-value also ranked high, underscoring the importance of temperature control and insulation in regulating energy usage. The type of building appeared consistently throughout the ranking, indicating its substantial influence on overall energy demand and usage patterns. The relevance of orientation and weather in photovoltaic power generation emphasizes the need to consider building direction for optimal energy production. These results provide valuable information for stakeholders to understand these critical features and design effective strategies aimed at reducing energy consumption, improving energy efficiency, and promoting sustainability in the built environment.

Finally, the prediction of each end-use demand is multiplied by its respective Irish primary energy factor, and these values are then summed to determine the total energy consumption of the building. This cumulative total is then divided by the area of the building to calculate the EUI, a measure of the energy performance of the building. The results illustrate the significant improvement in the performance of various machine learning models in predicting EUI with and without applying segregation methods (Fig. 8). Firstly, non-segregation scenario, the XGB model demonstrates the best performance on all metrics, boasting an RMSE of 13.89, MAE of 9.72, and an accuracy of 76% in terms of building rating. LGBM follows closely in performance. However, as we move down the table, the performance degrades, with the SVM having an RMSE of 71.96, MAE of 50.98, R-squared of 0.76 and accuracy of 29%. This suggests that the Gradient Boosts models, such as XGB and LGBM, are better suited for this problem of non-segregation.

Secondly, when considering the EUI Segregation scenario, there is a notable enhancement in the performance of several models. Specifically, the XGB and LGBM models excel with good R-squared values and substantially lower RMSE and MAE values compared to those without the segregation method. These models achieve substantially higher accuracy, with XGB reaching 89% and LGBM reaching 87%. This signifies that segregation could efficiently capture the underlying data patterns, aiding these models in making more precise predictions. However, it is essential to note that some models, such as NN, LR, KNN, and SVM, continue to demonstrate suboptimal performance even in the segregation scenario. The Neural Network (NN) model shows relatively less improvement compared to other models, which might suggest that it does not benefit as much from segregation in this particular context. The poor performance of SVM persisted even with segregation, indicating that this model might not be suitable for this dataset irrespective of the data processing method.

These results indicate that incorporating segregation in the analysis improves the performance of most models, particularly XGB, LGBM, and HGB. These findings highlight the importance of considering segregation in the machine learning process to obtain more accurate predictions for EUI values and emphasize the potential for future research to explore novel approaches to improve the performance of models that are lagging.

The modeling process is further improved using ensemble learning techniques to combine the best-developed models (XGB, LGBM, and HGB) based on performance. By comparing the interpretation of these models, this study seeks to identify the most effective approach for predicting building energy performance using machine learning techniques.

These results highlight the importance of EUI segregation and the effectiveness of ensemble modeling in improving the accuracy of end-use demand prediction (Table 6). In general, non-segregation method, the XGB model achieved an RMSE of 13.89, with an accuracy of 76%. On the contrary, the XGB model segregation method results in a significantly lower RMSE of 7.69, indicating reduced prediction errors compared to the previous method. The accuracy improves to 89%, suggesting more accurate predictions in most cases. Finally, the ensemble-based segregation approach, combining the XGB, LGBM, and HGB models, achieves the lowest RMSE of 6.48, demonstrating a further reduction in prediction errors compared to the previous methods. Accuracy reaches 91%, indicating a higher level of correct predictions than the other methods. The confusion matrix shows that the model performs well with all energy ratings of the building (Fig. 9). The findings suggest that the combination of models can enhance prediction capabilities and provide more reliable estimates for decision-making processes.

4.5. Urban building energy performance analysis

In the urban building energy performance analysis phase, the developed model is applied to practical application scenarios, implementing retrofit measures outlined in Ireland's National Climate Action Plan 2023. The objective is to retrofit existing residential buildings with below B2 ratings and install heat pumps. Two different scenarios are developed, improving the U values of windows, walls and roofs as recommended by Part L of the Irish Building Regulations and upgrading the HVAC system from a boiler to a heat pump. Additionally, the scenarios include options with and without renewables (Table 7).

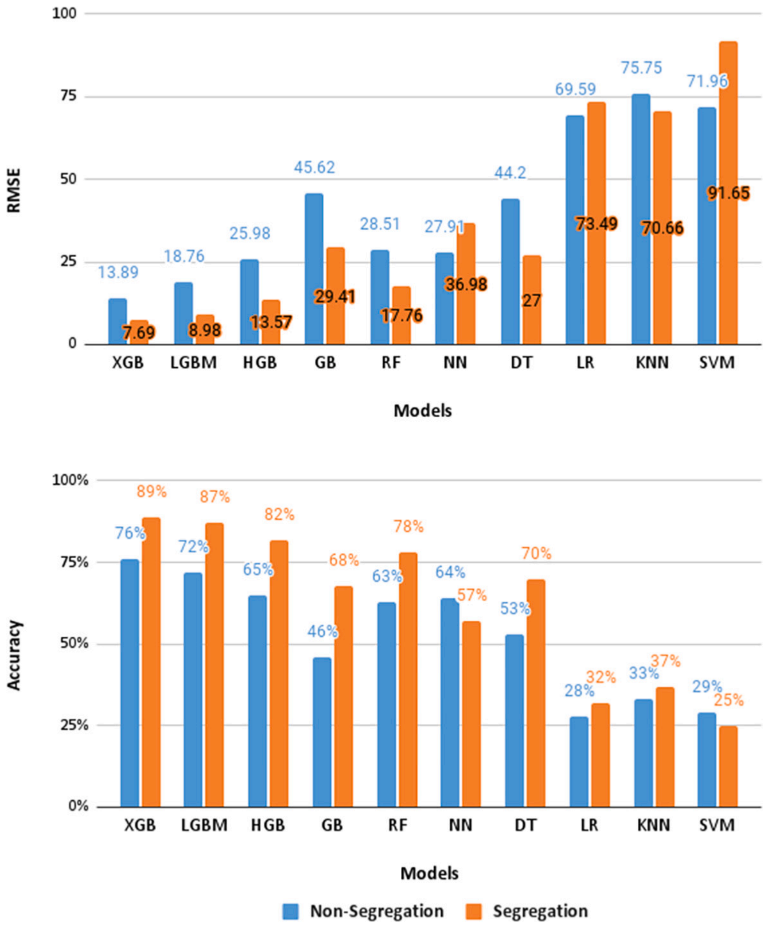


Fig. 8. Comparative analysis RMSE and accuracy of machine learning models using with and without end-use demand segregation method to predict EUI.

Table 6
Comparative analysis of method and machine learning models for predicting EUI using model performance metrics.

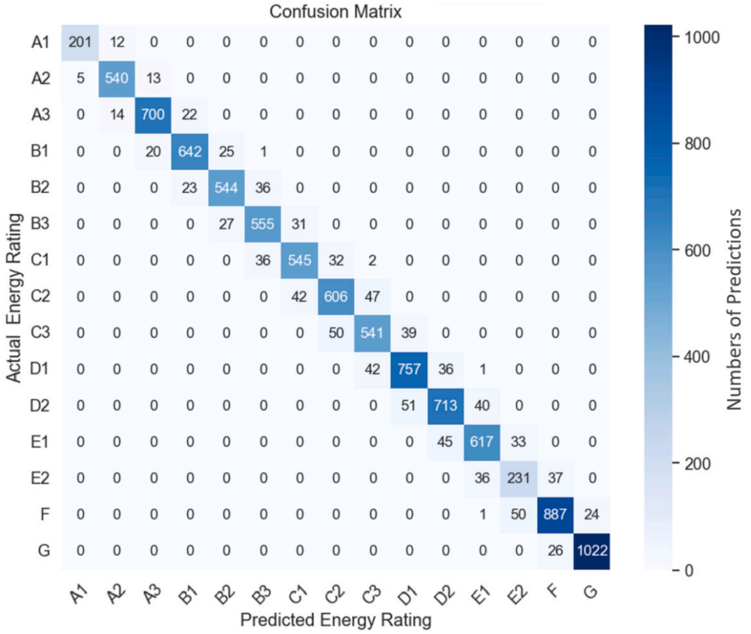
Methods	Models	RMSE	MAE	R-squared	Accuracy
Non-Segregation	XGB	13.89	9.72	0.99	76%
Segregation	XGB	7.69	4.67	1	89%
Ensemble Segregation	XGB, LGBM, HGB	6.48	3.9	1	91%

Both retrofit scenarios are applied to a dataset of 10,000 buildings with ratings below B2 and boilers as the HVAC system. This dataset size of 10,000 buildings allows for a sufficiently large sample to analyze and apply retrofit scenarios effectively, covering all inefficient building ratings from B3 to G. In general, there is a significant improvement in the distribution of energy ratings in buildings. Furthermore, implementing both retrofit scenarios in sample buildings resulted in a notable improvement, as indicated by the change in the distribution curve from lower energy ratings to higher ones (Fig. 10). However, the results indicate that in Scenario I, where the heat pumps are installed with windows, walls, and roofs refurbished, only 2,725 buildings achieved a rating of B2 and above.

In contrast, Scenario II, which included renewable installations, showed a slight improvement, with 3,467 buildings reaching higher ratings. These results demonstrate that both scenarios could only improve the higher rating of a relatively small percentage of buildings, ranging from 27% to 34%. It highlights the need for deeper retrofitting measures to achieve higher ratings, including heat pumps and renewables (Fig. 10).

The results are further examined using historical and future weather conditions, utilizing a 2030-year weather file. The emission scenarios considered in this study are based on a Representative Concentration Pathway (RCP), which is a greenhouse gas concentration trajectory adopted by the IPCC [60]. The 2030 weather file is based on RCP 4.5, described by the IPCC as an intermediate scenario and the most probable baseline scenario, considering the exhaustible nature of non-renewable fuels. The study shows no significant differences when using the future weather file. However, due to global warming and projected average temperature increases of 1–1.6 °C, heating demand is expected to decrease in the future, potentially leading to an improvement in building energy ratings [61]. Furthermore, the rating distribution for buildings is expected to change, primarily through using photovoltaics as renewable energy sources (Fig. 11).

The results demonstrate that the proposed methodology helps urban planners, energy policymakers, utility planners, and manufacturers in evaluating the implementation of retrofit measures on a large scale. Additionally, this case study highlights that fabric renovation in buildings is insufficient as a standalone solution. In conjunction with the installation of the heat pump, it is crucial to address other factors such as the airtightness of the building and the control of the heating to effectively improve the energy performance of the building, as evidenced by the importance of the characteristics.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	E1	E2	F	G
A1	201	12	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	5	540	13	0	0	0	0	0	0	0	0	0	0	0	0
A3	0	14	700	22	0	0	0	0	0	0	0	0	0	0	0
B1	0	0	20	642	25	1	0	0	0	0	0	0	0	0	0
B2	0	0	0	23	544	36	0	0	0	0	0	0	0	0	0
B3	0	0	0	0	27	555	31	0	0	0	0	0	0	0	0
C1	0	0	0	0	0	36	545	32	2	0	0	0	0	0	0
C2	0	0	0	0	0	0	42	606	47	0	0	0	0	0	0
C3	0	0	0	0	0	0	0	50	541	39	0	0	0	0	0
D1	0	0	0	0	0	0	0	0	42	757	36	1	0	0	0
D2	0	0	0	0	0	0	0	0	0	51	713	40	0	0	0
E1	0	0	0	0	0	0	0	0	0	0	45	617	33	0	0
E2	0	0	0	0	0	0	0	0	0	0	0	36	231	37	0
F	0	0	0	0	0	0	0	0	0	0	0	1	50	887	24
G	0	0	0	0	0	0	0	0	0	0	0	0	0	26	1022

Fig. 9. Confusion matrix shows the performance of the ensemble-based segregation model for each building rating. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 7
Retrofit scenarios to analyze the pre or post-effect on building energy performance at urban scale.

Retrofit Scenarios	Window U-value	Wall U-value	Roof U-value	HVAC	Renewables
Scenario I	1.4	0.21	0.16	Heat Pump	No
Scenario II	1.4	0.21	0.16	Heat Pump	Yes

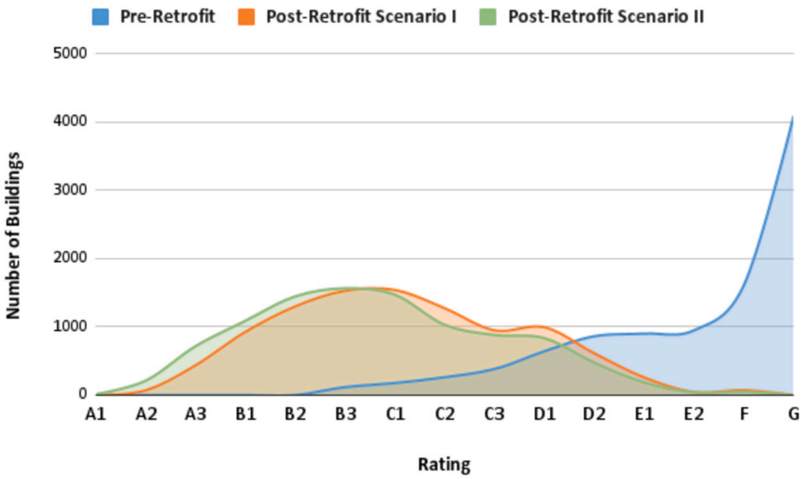


Fig. 10. Impact on the distribution of 10,000 building sample pre or post-retrofit scenarios.

5. Discussion

The proposed data-driven methodology offers a potential solution by enabling the analysis of the energy performance of residential buildings on a large scale, facilitating the decision-making process. The methodology uses limited available data to generate a synthetic dataset of 1 million buildings. This dataset is then used to develop a machine-learning model explicitly designed for the urban context. However, the data required to implement the proposed methodology, such as building geometry and non-geometry data, census information, and weather data, originate from various sources and come in different formats, leading to data inconsistencies. Consequently, due to these inconsistencies and the absence of standardized urban-scale data, available data present a significant and ongoing barrier to accurately implementing urban-scale modeling. The developed model allows for the prediction of various retrofit scenarios, even with limited resources. Segregation and ensemble-based methods improve the overall performance of the model, resulting in a significant 15% improvement. However, it is essential to note that the accuracy and implementation of the model depend on the quality and availability of input data and may vary in different contexts and countries. Moreover, developing synthetic data for different building archetypes in other contexts might require additional computational time.

Furthermore, the study identifies the key characteristics that influence the building demand for end-use. This finding enables policymakers to prioritize these influential features when considering retrofit

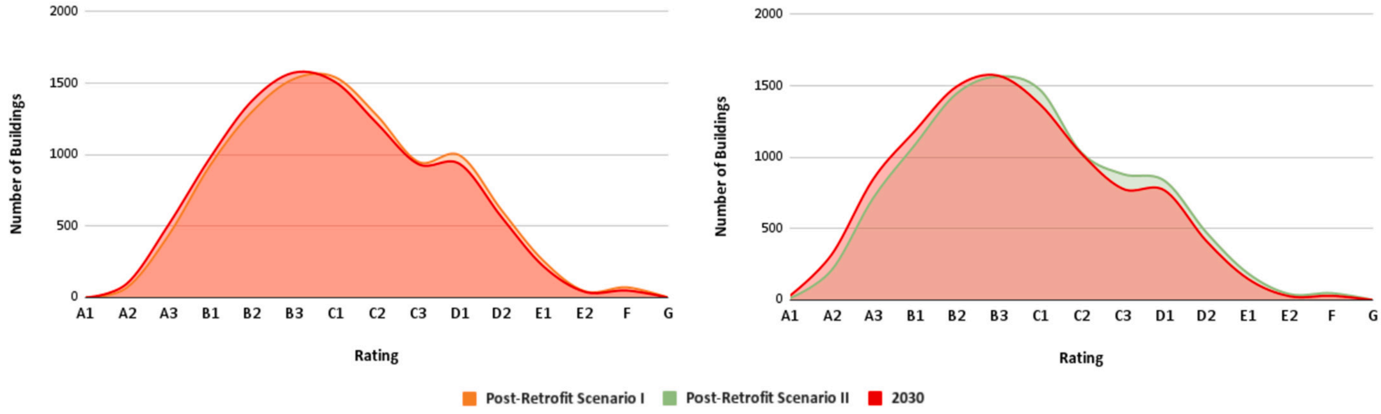


Fig. 11. Impact of historical and future weather conditions on the post-retrofit scenarios.

measures. By focusing on these critical factors, policymakers can effectively allocate resources and implement targeted retrofit strategies to improve building energy efficiency. However, it should be acknowledged that the importance of characteristics may differ for different sample data, weather conditions, or urban contexts.

Finally, the proposed solution is a valuable tool for urban planners, energy policymakers, utility planners, and manufacturers in evaluating and implementing retrofit scenarios at the urban scale. However, the models inherently depend on the quality of the data input. Therefore, incorrect synthetic data that do not closely represent real-world conditions might not accurately capture the complexities and uncertainties of the actual urban context. Furthermore, machine learning models are often considered ‘black boxes,’ which could lead to a lack of understanding of the underlying reasons behind the predictions. This lack of knowledge makes it difficult for policymakers and planners to trust and fully understand the recommendations. Additionally, the complexity and computational requirements of machine learning models and parametric simulations can be prohibitive, necessitating significant computational resources.

6. Conclusion and future work

Stakeholders analyze the energy performance of buildings on an urban scale to develop effective policy measures that reduce energy consumption and CO_2 emissions. However, collecting and analyzing building energy performance data on a large scale is complex and time-consuming, requiring multiple resources. To address this challenge, we propose a novel methodology that uses machine learning algorithms to predict the energy performance of an entire urban building stock. This methodology allows stakeholders to make informed decisions and implement targeted interventions to promote sustainable urban development. In this paper, we implement the end-use demand segregation method and the ensemble-based approach to develop a robust learning model to predict building energy performance. This approach improves the predictive performance of machine learning and supports informed decision-making in building energy performance assessment.

The methodology tested on Dublin City by developing a synthetic building dataset of 1 million residential buildings using parametric analysis of 19 key parameters identified from four building archetypes. The results show that the segregation method is highly effective for predicting EUI based on the given dataset, compared to the traditional single model approach. Among the ten different machine learning algorithms compared, variations of the Gradient Boosting algorithm (XGB, LGBM, and HGB) are found to be the most efficient and accurate models to predict building energy performance. Furthermore, the ensemble-based approach further improved the results, achieving an accuracy of 91%. Comparing the ten different models revealed that the ensemble-based segregation method is highly effective in predicting EUI, with an improvement in the energy rating of the building resulting in an increase in accuracy 15%. Accurate prediction of building energy performance enables stakeholders, such as energy policymakers and urban planners, to make informed decisions when planning large-scale retrofit measures.

In general, the proposed methodology offers valuable information and tools to support urban planners and energy policymakers in addressing the challenges of sustainable planning and energy efficiency on an urban scale. The data-driven approach, coupled with feature analysis and predictive modeling, empowers decision-makers to make informed choices and drive positive change in urban energy systems. The findings of this study offer valuable assistance to energy policymakers and urban planners by providing information that can contribute to the development of effective retrofit measures. These measures aim to decrease building energy consumption and mitigate carbon emissions. By incorporating the knowledge gained from this study, policymakers and planners can make well-informed decisions that facilitate sustainable urban development and address the pressing issue of climate change. Furthermore, the study helps policymakers and urban planners evaluate the feasibility and impact of implementing retrofit measures on a larger scale. This comprehensive approach supports the formulation and execution of strategies to address energy efficiency and environmental concerns.

Future research directions could investigate the influence of different mid-rise or high-rise apartments and non-residential archetype models on the predictive performance of machine learning algorithms. Furthermore, the integration of cloud computing parametric simulation could further enhance the research results. Currently, this research focuses on annual energy use and could be expanded to analyze seasonal and monthly variations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This publication has emanated from research supported by Science Foundation Ireland through US-Ireland R&D Partnership Research Grant 20/US/3695, the U.S. National Science Foundation through Award Number 2217410, and the Department for the Economy in Northern Ireland through USI 167. The opinions, findings, and conclusions or recommendations expressed in this material are those of the

author(s) and do not necessarily reflect the views of the Science Foundation Ireland or other funding agencies.

References

[1] EU-Energy, Energy for Europe by European commission, Online; https://energy.ec.europa.eu/index_en, 2022. (Accessed 1 December 2022).

[2] W.A. Benjamin, Revision of the energy performance of buildings directive: fit for 55 package, 2022.

[3] U. Ali, M.H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, Appl. Energy 267 (2020) 114861.

[4] C. Hoare, R. Aghamolaei, M. Lynch, A. Gaur, J. O'Donnell, A linked data approach to multi-scale energy modelling, Adv. Eng. Inform. 54 (2022) 101719.

[5] C.F. Reinhart, C.C. Davila, Urban building energy modeling–a review of a nascent field, Build. Environ. 97 (2016) 196–202.

[6] T. Hong, Y. Chen, X. Luo, N. Luo, S.H. Lee, Ten questions on urban building energy modeling, Build. Environ. 168 (2020) 106508.

[7] U. Ali, M.H. Shamsi, C. Hoare, E. Mangina, J. O'Donnell, Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis, Energy Build. 246 (2021) 111073.

[8] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review, Energy Build. 165 (2018) 301–320.

[9] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: load prediction, pattern identification, fault detection and diagnosis, Energy Built Environ. 1 (2) (2020) 149–164.

[10] Y. Wang, T. Wu, H. Li, M. Skitmore, B. Su, A statistics-based method to quantify residential energy consumption and stock at the city level in China: the case of the Guangdong-Hong Kong-Macao Greater Bay area cities, J. Clean. Prod. 251 (2020) 119637.

[11] Y. Sun, F. Haghighat, B.C. Fung, A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energy Build. 221 (2020) 110022.

[12] C. Tian, Y. Ye, Y. Lou, W. Zuo, G. Zhang, C. Li, Daily power demand prediction for buildings at a large scale using a hybrid of physics-based model and generative adversarial network, in: Building Simulation, vol. 15, Springer, 2022, pp. 1685–1701.

[13] N. Abbasabadi, M. Ashayeri, Urban energy use modeling methods and tools; a review and an outlook for future tools, Build. Environ. (2019) 106270.

[14] P. Manandhar, H. Rafiq, E. Rodriguez-Ubinas, Current status, challenges, and prospects of data-driven urban energy modeling: a review of machine learning methods, Energy Rep. 9 (2023) 2757–2776.

[15] C. Benavente-Peces, N. Ibadah, Buildings energy efficiency analysis and classification using various machine learning technique classifiers, Energies 13 (13) (2020) 3497.

[16] U. Ali, M.H. Shamsi, F. Alshehri, E. Mangina, J. O'Donnell, Comparative analysis of machine learning algorithms for building archetypes development inurban building energy modeling, in: Building Performance Modeling Conference and SimBuild, 2018.

[17] Y. Chen, M. Guo, Z. Chen, Z. Chen, Y. Ji, Physical energy and data-driven models in building energy prediction: a review, Energy Rep. 8 (2022) 2656–2671.

[18] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, S. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022) 103406.

[19] Y. Pan, M. Zhu, Y. Lv, Y. Yang, Y. Liang, R. Yin, Y. Yang, X. Jia, X. Wang, F. Zeng, et al., Building energy simulation and its application for building performance optimization: a review of methods, tools, and case studies, Adv. Appl. Energy (2023) 100135.

[20] M. Ferrando, F. Causone, T. Hong, Y. Chen, Urban building energy modeling (UBEM) tools: a state-of-the-art review of bottom-up physics-based approaches, Sustain. Cities Soc. 62 (2020) 102408.

[21] O. Pasichnyi, J. Wallin, O. Kordas, Data-driven building archetypes for urban building energy modelling, Energy 181 (2019) 360–377.

[22] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (8) (2009) 1819–1835.

[23] T. Hong, Y. Chen, S.H. Lee, M.A. Piette, Citybes: a web-based platform to support city-scale building energy efficiency, Urban Comput. 14 (2016) 2016.

[24] Y. Chen, T. Hong, M.A. Piette, Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis, Appl. Energy 205 (2017) 323–335.

[25] D. Robinson, F. Haldi, P. Leroux, D. Perez, A. Rasheed, U. Wilke, Citysim: comprehensive micro-simulation of resource flows for sustainable urban planning, in: Proceedings of the Eleventh International IBPSA Conference, no. CONF, 2009, pp. 1083–1090.

[26] C. Reinhart, T. Dogan, J.A. Jakubiec, T. Rakha, A. Sang, Umi-an urban simulation environment for building energy use, daylighting and walkability, in: 13th Conference of International Building Performance Simulation Association, Chambery, France, 2013.

[27] C.C. Davila, C.F. Reinhart, J.L. Bemis, Modeling Boston: a workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets, Energy 117 (2016) 237–250.

[28] R. El Kontar, B. Polly, T. Charan, K. Fleming, N. Moore, N. Long, D. Goldwasser, Urbanopt: an open-source software development kit for community and urban district energy modeling, Tech. rep., National Renewable Energy Lab. (NREL), Golden, CO (United States), 2020.

[29] Y.Q. Ang, Z.M. Berzolla, S. Letellier-Duchesne, V. Jusiega, C. Reinhart, Ubem. io: a web-based framework to rapidly generate urban building energy models for carbon reduction technology pathways, Sustain. Cities Soc. 77 (2022) 103534.

[30] S.S. Abolhassani, M. Amayri, N. Bouguila, U. Eicker, A new workflow for detailed urban scale building energy modeling using spatial joining of attributes for archetype selection, J. Build. Eng. 46 (2022) 103661.

[31] A. Katal, M. Mortezazadeh, L.L. Wang, H. Yu, Urban building energy and microclimate modeling–from 3d city generation to dynamic simulations, Energy 251 (2022) 123817.

[32] I.M. Borràs, D. Neves, R. Gomes, Using urban building energy modeling data to assess energy communities' potential, Energy Build. 282 (2023) 112791.

[33] A. Nutkiewicz, Z. Yang, R.K. Jain, Data-driven urban energy simulation (due-s): integrating machine learning into an urban building energy simulation workflow, Energy Proc. 142 (2017) 2114–2119.

[34] A. Rahman, V. Srikumar, A.D. Smith, Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks, Appl. Energy 212 (2018) 372–385.

[35] C.E. Kontokosta, C. Tull, A data-driven predictive model of city-scale energy use in buildings, Appl. Energy 197 (2017) 303–317.

[36] F. Jiang, J. Ma, Z. Li, Y. Ding, Prediction of energy use intensity of urban buildings using the semi-supervised deep learning model, Energy 249 (2022) 123631.

[37] Y. Zhang, B.K. Teoh, M. Wu, J. Chen, L. Zhang, Data-driven estimation of building energy consumption and ghg emissions using explainable artificial intelligence, Energy 262 (2023) 125468.

[38] J. Seo, S. Kim, S. Lee, H. Jeong, T. Kim, J. Kim, Data-driven approach to predicting the energy performance of residential buildings using minimal input data, Build. Environ. 214 (2022) 108911.

[39] N.-T. Ngo, A.-D. Pham, T.T.H. Truong, N.-S. Truong, N.-T. Huynh, T.M. Pham, An ensemble machine learning model for enhancing the prediction accuracy of energy consumption in buildings, Arab. J. Sci. Eng. 47 (4) (2022) 4105–4117.

[40] M. Wurm, A. Droin, T. Stark, C. Geiß, W. Sulzer, H. Taubenböck, Deep learning-based generation of building stock data from remote sensing for urban heat demand modeling, ISPRS Int.l J. Geo-Inf. 10 (1) (2021) 23.

[41] A.S. Mohammed, P.G. Asteris, M. Koopialipoor, D.E. Alexakis, M.E. Lemonis, D.J. Armaghani, Stacking ensemble tree models to predict energy performance in residential buildings, Sustainability 13 (15) (2021) 8298.

[42] F. Johari, G. Peronato, P. Sadeghian, X. Zhao, J. Widén, Urban building energy modeling: state of the art and future prospects, Renew. Sustain. Energy Rev. 128 (2020) 109902.

[43] T. Loga, B. Stein, N. Diefenbach, Tabula building typologies in 20 European countries—making energy-related features of residential building stocks comparable, Energy Build. 132 (2016) 4–12.

[44] U. Ali, M.H. Shamsi, C. Hoare, E. Mangina, J. O'Donnell, A data-driven approach for multi-scale building archetypes development, Energy Build. 202 (2019) 109364.

[45] W. Wang, S. Li, S. Guo, M. Ma, S. Feng, L. Bao, Benchmarking urban local weather with long-term monitoring compared with weather datasets from climate station and energyplus weather (EPW) data, Energy Rep. 7 (2021) 6501–6514.

[46] M.P. Tootkaboni, I. Ballarini, M. Zinzi, V. Corrado, A comparative analysis of different future weather data for building energy performance simulation, Climate 9 (2) (2021) 37.

[47] Y. Zhang, I. Korolija, Performing complex parametric simulations with jeplus, in: SET2010-9th International Conference on Sustainable Energy Technologies, 2010, pp. 24–27.

[48] J. Egan, D. Finn, P.H.D. Soares, V.A.R. Baumann, R. Aghamolaei, P. Beagon, O. Neu, F. Pallonetto, J. O'Donnell, Definition of a useful minimal-set of accurately-specified input data for building energy performance simulation, Energy Build. 165 (2018) 172–183.

[49] Y. Choi, D. Song, S. Yoon, J. Koo, Comparison of factorial and Latin hypercube sampling designs for meta-models of building heating and cooling loads, Energies 14 (2) (2021) 512.

[50] W. Tian, Y. Heo, P. De Wilde, Z. Li, D. Yan, C.S. Park, X. Feng, G. Augenbroe, A review of uncertainty analysis in building energy assessment, Renew. Sustain. Energy Rev. 93 (2018) 285–301.

[51] Y. Ye, M. Strong, Y. Lou, C.A. Faulkner, W. Zuo, S. Upadhyaya, Evaluating performance of different generative adversarial networks for large-scale building power demand prediction, Energy Build. 269 (2022) 112247.

[52] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.

[53] Building energy rating certificate database by SEAI, Online; https://ndber.seai.ie/BERResearchTool/ber/search.aspx. (Accessed 25 October 2023).

[54] K. McDonagh, Geodirectory technical guide, an post and ordnance survey Ireland, 2023.

[55] Ordnance survey Ireland, Online; https://www.osi.ie. (Accessed 25 October 2023).

[56] Census of population 2022 - profile 1 housing in Ireland by Central Statistics Office, Online; https://www.cso.ie/en/releasesandpublications/ep/p-cpsr/censusofpopulation2022-summaryresults/, 2022. (Accessed 25 October 2023).

[57] Ireland climate action plan 2023, Online; https://www.gov.ie/en/publication/7bd8c-climate-action-plan-2023/. (Accessed 25 October 2023).

[58] U. Ali, M.H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, Appl. Energy 267 (2020) 114861.

[59] J. Laue, Ashrae 62.1: using the ventilation rate procedure, Consult.-Specif. Eng. 55 (2018) 14–17.

[60] Intergovernmental panel on climate change (IPCC), Online; https://www.ipcc.ch. (Accessed 25 October 2023).

[61] P. Nolan, J. Flanagan, High-resolution climate projections for Ireland–a multi-model ensemble approach, Environmental Protection Agency, 2020.

[62] D. Sood, I. Alhindawi, U. Ali, J.A. McGrath, M.A. Byrne, D. Finn, J. O'Donnell, Simulation-based evaluation of occupancy on energy consumption of multi-scale residential building archetypes, J. Build. Eng. 75 (2023) 106872.

[63] D. Sood, I. Alhindawi, U. Ali, D. Finn, J.A. McGrath, M.A. Byrne, J. O'Donnell, Zone-wise occupancy schedules developed using Time Use Survey data for building energy performance simulations, Data Brief 49 (2023) 109453.