| **Title** | A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings |
|---|---|
| **Authors(s)** | Ali, Usman; Shamsi, Mohammad Haris; Bohacek, Mark; Hoare, Cathal; Purcell, Karl; Mangina, Eleni; O'Donnell, James |
| **Publication date** | 2020-01-01 |
| **Publication information** | Applied Energy, 267 : , Article Number: 114861 |
| **Publisher** | Elsevier |
| **Item record/more information** | http://hdl.handle.net/10197/12266 |
| **Publisher's statement** | This is the author's version of a work that was accepted for publication in Applied Energy. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Applied Energy (267 (2020)) https://doi.org/10.1016/j.apenergy.2020.114861 |
| **Publisher's version (DOI)** | 10.1016/j.apenergy.2020.114861 |
A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings

Usman Ali\textsuperscript{a}, Mohammad Haris Shamsi\textsuperscript{a}, Mark Bohacek\textsuperscript{b}, Cathal Hoare\textsuperscript{a}, Karl Purcell\textsuperscript{b}, Eleni Mangina\textsuperscript{c}, James O’Donnell\textsuperscript{a}

\textsuperscript{a}School of Mechanical and Materials Engineering and UCD Energy Institute, UCD, Dublin, Ireland
\textsuperscript{b}Sustainable Energy Authority Of Ireland, Dublin, Ireland
\textsuperscript{c}School of Computer Science and UCD Energy Institute, UCD, Dublin, Ireland

Abstract

Urban planners face significant challenges when identifying building energy efficiency opportunities and developing strategies to achieve efficient and sustainable urban environments. A possible scalable solution to tackle this problem is through the analysis of building stock databases. Such databases can support and assist with building energy benchmarking and potential retrofit performance analysis. However, developing a building stock database is a time-intensive modeling procedure that requires extensive data (both geometric and non-geometric). Furthermore, the available data for developing a building database is sparse, inconsistent, diverse and heterogeneous in nature. The main aim of this study is to develop a generic methodology to optimize urban scale energy retrofit decisions for residential buildings using data-driven approaches. Furthermore, data-driven approaches identify the key features influencing building energy performance. The proposed methodology formulates retrofit solutions and identifies optimal features for the residential building stock of Dublin. Results signify the importance of data-driven retrofit modeling as the feature selection process reduces the number of features in Dublin’s building stock database from 203 to 56 with a building rating prediction accuracy of 86%. Amongst the 56 features, 16 are identified to be recommended as retrofit measures (such as fabric renovation values and heating system upgrade features) associated with each energy-efficiency rating. Urban planners and energy policymakers could use this methodology to optimize large-scale retrofit implementation, particularly at an urban scale with limited resources. Furthermore, stakeholders at the local authority level can estimate the required retrofit investment costs, emission reductions and energy savings using the target retrofit features of energy-efficiency ratings.

Keywords: building energy retrofit, building energy performance, urban building energy analysis, building energy stock, retrofit modeling, data-driven approach
Nomenclature

$\tilde{c}_i$ target value

$ACC$ ACCuracy

$AERG$ Advanced Energy Retrofit Guide

$ASIEPI$ ASessment and Improvement of the EPBD Impact

$BEH$ Better Energy Homes

$BER$ Building Energy Rating

$BEWH$ Better Energy Warmer Homes

$BPD$ Buildings Performance Database

$BPIE$ Buildings Performance Institute Europe

$c_i$ predicted value

$CBECS$ Commercial Building Energy Consumption Survey

$CE$ Classification Error

$CSO$ Central Statistics Office

$DEAP$ Dwelling Energy Assessment Procedure

$ENTRANZE$ ENforce the TRANsition to Nearly Zero Energy buildings

$EPBD$ European Union Energy Performance of Buildings Directive

$EPC$ Energy Performance Certificate

$FN$ number of false negatives
1. Introduction

Buildings account for almost 36% of global CO₂ emissions and approximately one-third of energy consumption [1]. Since 2010, building related CO₂ emissions have increased by nearly 1% annually [2]. One key reason is the inefficient energy performance of the building stock while attempting to deliver intended functional intent for the spaces within. In Europe, more than 35% of the buildings are more than 50 years old and only 0.4 - 1.2% of buildings are retrofitted per year (depending on the country) [1]. Multiple economic, social and environmental benefits arise from energy-related improvements to buildings. Consequently, different energy policies have been crafted over the past few years. In Europe, the 2010 Energy Performance of Buildings Directive 2010/31/EU and the 2012 Energy Efficiency Directive aim to improve the energy performance of buildings within the European Union members states [3]. This directive received a major update in 2018 (2018/844/EU) with the focus shifting towards building renovation and smart buildings [4].

Improving buildings’ energy performance is an essential step in the reduction of energy and emissions related to building stock [5]. A possible scalable solution to tackle this problem

| Term   | Definition                                      |
|--------|-------------------------------------------------|
| FP     | number of false positives                       |
| GBPN   | Global Buildings Performance Network            |
| K      | training or testing samples                     |
| $kNN$  | $k$-Nearest-Neighbours                          |
| $lrd$  | local reachability density                      |
| LVQ    | Learning Vector Quantization                    |
| $MinPts$ | Minimum Points                                  |
| $N$    | number of negatives in ground truth            |
| NZEB   | Nearly Zero Energy Buildings                    |
| $P$    | number of positives in ground truth             |
| RECS   | Residential Energy Consumption Survey           |
| RMSE   | Root Mean Squared Error                         |
| SEAI   | Sustainable Energy Authority of Ireland         |
| TABULA | Typology Approach for Building Stock Energy Assessment |
| $TP$   | number of true positives                        |
| UBEM   | Urban Building Energy Modeling                  |
is through the analysis of a building stock database that contains existing and new buildings. Such building stock databases inform policy formulation, planning, decision making, and analysis. Generally, building energy performance assessment relies on datasets that can support and assist activities such as energy profiling, building energy benchmarking, and detailed retrofit analysis. These building stock datasets provide additional information about buildings, including usage patterns, technical characteristics, and fuel consumption. Urban building stock modelers often use this information for energy analysis and modeling.

A building stock is broadly characterized into residential and non-residential buildings. Common examples of residential building stocks include houses and apartments. Non-residential building stocks include industrial and commercial complexes. Census and survey data are the two most crucial resources for gathering information about building stock. Generally, government or research institutes also collect information via census, questionnaire and large scale surveys of buildings; examples of such surveys are Residential Energy Consumption Survey (RECS), Commercial Building Energy Consumption Survey (CBECS), and Energy Performance Certificates (EPCs).

Urban Building Energy Modeling (UBEM) are often used to quantify residential buildings and develop the building stock database at a national level. UBEMs make use of physical properties of individual buildings and employ building energy simulation tools to model and simulate the use of a single building archetype, thus, representing the actual building stock. In the event of data being limited or almost non-existent, data-driven urban energy modeling could be used to quantify the urban building stock and predict the energy rating.

The development of a database for an entire building stock is a time-intensive modeling procedure and requires extensive geometric and non-geometric data. Furthermore, available data for developing a building database is typically sparse, inconsistent, diverse and heterogeneous in nature. The big challenges for stakeholders using these datasets are reliability, completeness, accuracy and consistency of the data. To date, a limited number of countries have developed a building stock database in the form of Energy Performance Certificates (EPCs) but these databases do not represent the entire building stock in these countries. Stakeholders often require entire building stock data to calculate accurate energy performance of buildings at a large scale. Furthermore, another important challenge of using these databases is to identify intelligent retrofit recommendations that can improve building energy performance.

In the European Energy Roadmap to 2050, the key goal is to retrofit and transform the existing stock into Nearly Zero Energy Buildings (NZEBs) by using renewable resources. Over the past few years, research conducted on building retrofitting mostly focuses on individual buildings. A significant challenge is to optimize and implement cost-effective retrofit recommendations for buildings at a large scale. Therefore, a consolidated knowledge base is required to help decision makers identify retrofit alternatives.

The main aim of this study is to develop a generalized methodology to optimize urban scale energy retrofit decisions for residential buildings using data-driven approaches. Furthermore, the devised data-driven methodology identifies key variables that influence building energy performance and help identify cost-effective retrofit recommendations. This study could be used to develop a consolidated knowledge base for residential building stock.
The resulting methodology will be used to furnish the already existing residential building stock database with information regarding the retrofit recommendation, retrofit implementation priorities and associated retrofit costs.

This paper is structured as follows: Section 2 describes an overview of existing work done in this domain and identifies the key research gaps; Section 3 describes a novel methodology, including an explanation of the different steps followed during knowledge base development; Section 4 evaluates the proposed methodology using an Irish case study and the results are discussed. Section 5 concludes this research study by describing challenges and future work.

2. Literature Review

2.1. Existing Building Stock Databases

Building stocks are broadly characterized into residential and non-residential buildings and data are gathered through census and survey data. Census data includes statistical building stock data at various scales (local, national, and international) while survey data involves additional sampling studies of individual buildings within a defined population of buildings.

The methodologies used to develop building stock databases vary by country. For instance, the United States’ Department of Energy maintains one of the largest building stock databases, the Building Performance Database (BPD), that contains information about the measured energy performance of residential and commercial building stock [14]. "Census Hub", a tool developed by the European Statistical System, provides access to the various national census databases of Europe [15]. The Building Performance Institute Europe (BPIE) also conducted an extensive survey to gather information about existing building stocks across Europe. Following the survey, BPIE released a Data Hub portal to report on the statistical data of the EU building stock [16]. Another initiative, ASsessment and Improvement of the EPBD Impact (ASIEPI), identified the potential problems when implementing EPBD in the EU [17]. Owing to these exhaustive measures, each member state in the EU maintains its own EPC database containing essential information about its building stock [18].

Generally, the local, urban, and national authorities employ EPC data to improve energy planning and decision making [19]. EPC data can also be used by energy consultancies, construction, and utility companies to provide different building services such as building renovation [20]. EPC data contains useful information; often needed to improve current building energy performance so as to target sustainable energy and climate goals [8]. Data quality associated with EPCs is a major concern as building energy rating assessors manually collect all the data. Each country measures data quality and mandates that assessors follow defined standards to calculate the final EPC of each building. However, the process is prone to human error and a significant number of opportunities exits to improve existing data [8].

Although the aforementioned studies provide a valuable overview of the existing building stock, there are several challenges associated with these databases that include: (1) a lack of physical descriptions for buildings, (2) often dated information (3) a lack of data quality, and (4) a lack of existing and suggested retrofit information [21] [22].
2.2. Existing Data-Driven Approaches for Building Stock Modeling

The quality of aforementioned databases is often debated as information is mostly gathered through surveys that are prone to human error, incomplete information, and irregular coverage sampling. As such, extracting useful, actionable and interesting knowledge from these databases often becomes a significant challenge for relevant stakeholders. Data-driven approaches provide an efficient solution for extracting useful knowledge from the raw data. Over the past decades, data-driven modeling has been extensively employed to model the existing building stock. For instance, data-driven methods are coupled with engineering simulation methods for predicting and modeling the energy performance of urban buildings. Similarly, machine learning techniques have been extensively used for building operational energy use modeling. Data-driven modeling is the optimal choice for prediction and classification of building energy consumption. Furthermore, this approach has been frequently implemented with urban scale applications in the energy modeling domain that include load forecasting, energy prediction, and energy pattern profiling. It is, therefore necessary to identify the opportunities offered by data-driven modeling for data quality improvement and knowledge extraction.

Building stock data consists of numerous features that describe individual buildings. Not all of these variables influence building energy performance to the same degree and relevant features must be identified. A feature selection process reduces the dimensionality of model inputs resulting in significantly lower computational loads. This process also enhances the performance of data-driven approaches by selecting the most appropriate inputs and reducing the redundancy within historical datasets. Feature selection generally uses engineering or data-driven techniques: Engineering methods leverage the analysis of experts’ interpretations and existing practices in literature; data-driven techniques, on the other hand, make use of statistical and data mining methods, for instance, regression analysis and neural networks.

A limited number of studies exist in the literature that focus on techniques that identify the most important features required to enable simplified energy modeling of residential buildings. For instances, relevant features can be identified using energy simulation with statistical analysis. Some of these variables include the building area, U-values, fuel type, dwelling type, and age band. Furthermore, the identified key factors help in determining the energy rating of existing residential building stock. The systematic feature selection procedure also helps in commercial building energy forecasting. However, most feature selection methods applied in the building energy domain focus on individual buildings and offer limited solutions for specific scenarios. There is an opportunity to develop generalized or automated solutions at urban and regional scale. Most studies use traditional static approaches as opposed to using data-driven features at an urban scale.

2.3. Existing Approaches for Urban Retrofit Modeling

Retrofitting is considered to be one of the effective approaches for reducing global energy consumption and greenhouse gas emissions. Despite extensive research in the retrofit domain, identifying cost-effective retrofit recommendations is still a challenging task for stakeholders. A building energy retrofit typically takes one of three forms, namely:
energy audit with no-cost changes (change of tariff structure or operational schedule), shallow retrofit, and deep energy retrofit [35]. Energy audits involve thorough reviews of energy utility bills and a preliminary analysis of building energy usage. This method usually results in average savings of less than 25% and involves minimal risk and investment [36]. Shallow retrofits yield quick, straightforward energy savings where individual retrofit measures apply to different building components such as the envelope, lighting, and heating control. This method normally results in 25 to 45% energy savings and involves little risk and limited costs [37]. For deep retrofits, refurbishments often include whole building components coupled with detailed analysis of the building. This approach typically results in energy savings of more than 45% [38]. Such savings are significantly higher than other methods as deep retrofitting employs detailed building analysis and experts’ opinions [39].

Recent studies mostly focus on retrofit modeling of individual or a group of buildings [40]. For instance, building energy performances assessment and Geographic Information System (GIS) are used to implement energy retrofit policies for residential buildings at the urban scale [41]. Due to privacy issues and limited data availability, not many studies focus on residential stock at a large scale. A limited number of studies examine the urban area but from the perspective of commercial buildings. For instance, City Building Energy Saver (CityBES) is a retrofit tool to analyze the potential retrofit energy use and energy cost savings for the offices and retail buildings at city-scale using generation and simulation of building energy models [42]. Similarly, simulation and optimisation methodologies are used to implement energy retrofits measures in commercial buildings [43]. The scope of these studies is often limited to a specific climate or certain pre-defined scenarios [44]. A generalized solution would enhance the scope and scale of urban scale retrofitting for residential building stock [45].

Accordingly, the Advanced Energy Retrofit Guide (AERG) by the U.S. Department of Energy suggests using all three types (energy audit, shallow retrofit, and deep energy retrofit) to achieve sustainable buildings [37]. A significant number of previous studies establish that shallow and deep retrofit procedures are the best solutions for achieving significant energy savings [12]. The deep retrofit procedure is recommended internationally [46]. This retrofit procedure has often been applied to individual buildings but large scale retrofitting methodologies are almost non-existent; this poses a big challenge for urban planners and energy policymakers [35]. It is often difficult to include different classes of buildings that exist at the urban scale. Therefore, there is a need for a retrofit knowledge base that stores the retrofit recommendations for different classes of buildings.

3. Methodology

Building stocks that possess energy performance data are increasingly being used for detailed energy analysis. Available data often originates from different sources and is usually unstructured in nature. Hence, the use of existing energy performance data gives rise to technological and methodological challenges for urban planners and energy policy makers [47]. The quality of building stock data has a significant impact on the accuracy of energy studies. Therefore, this paper proposes a generalized approach for the development
of a building energy performance knowledge base that is specifically tailored to residential
buildings. This knowledge base will also aid in identification of retrofit recommendations
for similar classes of buildings. The approach uses data-driven techniques to improve the
quality of available data and extract knowledge that can be used to significantly reduce
building energy consumption.

The devised methodology establishes a procedure for developing a knowledge base that
extracts useful information about a building from different sources. The developed knowl-
dge base would aid in the prediction of building energy performance. Furthermore, the de-
vised methodology provides a procedure to identify key retrofit recommendations for urban
scale residential buildings. The methodology also implements detailed, in-depth concepts
for retrofit modeling at the urban scale.

The development of a building energy performance knowledge base follows five steps
as shown in Figure 1. The initial step involves data collection from an existing building
stock. The pre-processing step follows and employs data-driven approaches to improve
the quality of the building stock data. The next step, feature selection, determines the
key features that influence retrofit choices for an existing building. The penultimate step,
urban retrofit modeling, determines cost-effective retrofit recommendations for the developed
building stock. Finally, the last step organizes a knowledge base that associates building
energy performance data with appropriate retrofit recommendations. The following sections
describe the individual steps of the methodology in further detail.

3.1. Data Collection

Data collection is the initial and the most crucial step in the knowledge base development
process. The goal is to collect and merge data from different sources.

| Data Type                  | Source                                      | Example                        |
|----------------------------|---------------------------------------------|--------------------------------|
| Geometric                  | National building stock or EPC database [14, 11] | Building shape, envelope      |
| Non-geometric              | National building stock or EPC database [14, 11] | Construction assemblies, HVAC systems |
| Building energy            | EPC database [11, 8, 17, 48]                | Energy use intensity           |
| Statistics                 | Census database [15]                        | Number of buildings            |
| Retrofit Measure Cost      | Construction companies or national retrofit scheme [49] | Building fabric renovation    |
| Retrofit Expert Reports    | National Housing Guideline [50, 51]         | Cost optimal reports           |

At the urban scale, information regarding the building stock is often extracted with the
help of existing building data for the geographical area under consideration. Building data
contains geometric, non-geometric, and energy performance index information. Geometric
Figure 1: Overarching data-driven methodology for residential building energy performance knowledge base development and determination of key retrofit recommendations at urban scale
data includes building shape, dwelling type, building envelope, number of floors, walls, and windows. On the other hand, non-geometric building data includes envelope U-values, construction assemblies, and HVAC systems. Data for building energy performance indices include energy use intensity information, which can be acquired from EPC databases. Data collection also requires quantification of buildings by type in a given geographical area. Generally, census data (national statistics) contains information regarding the number of buildings at country or regional level.

Data collection gathers existing energy reports and experts’ opinions regarding building retrofits. This information is crucial and required for the retrofit modeling process (Section 3.4). Furthermore, the data collection process also aggregates retrofit implementation costs to conduct an economic analysis (Table 1).

3.2. Building Stock Pre-processing

This process follows data collection and involves an initial statistical analysis, data pre-processing, outlier detection and building archetypes development procedures (Figure 2). Furthermore, this process employs various data-driven techniques such as data pre-processing and outlier detection to extract building characteristics and associated energy usage of individual buildings.

Figure 2: The pre-processing workflow of the building stock data to enhance the quality of building characteristics information

This study relies on EPC data which is publicly available in most countries for building stock development and enhancement. As the EPC data are usually collected from statistical surveys or questionnaires, many anomalies can exist in the dataset. Data-driven methods such as data pre-processing and outlier detection can eliminate/treat these existing anomalies and aid in the extraction of clean data.

The first sub-step involves a crude statistical analysis of the building stock dataset. Visual analysis of statistical representations (that include density plots, histograms and box plots)
are used to examine the quality of data and identify appropriate processing techniques. Real-world data and data collected through surveys such as EPCs, generally contain irrelevant, incomplete, noisy, redundant, and inconsistent information; this makes it challenging to extract useful or accurate knowledge. Data pre-processing follows statistical analysis and eliminates identified inconsistencies before the data are further used. Pre-processing uses the following steps [52]:

- **Data cleaning**: fills in missing or zero values, removes duplicate data, smooths out noisy data, and resolves inconsistencies. An example would be to replace missing or zero building fabric area values with an average or median value.

- **Data integration**: combines data from multiple sources such as redefinition of building type description according to a common standard datatype.

- **Data reduction**: removes irrelevant attributes such as IDs, date and time.

- **Data transformation**: replaces or adds new variables inferred from existing variables such as overall building U-value.

- **Data discretization**: replaces numerical attributes with nominal values, for example, convert from year of construction into age bands.

The third sub-step, outlier detection, follows data pre-processing and further enhances the quality of data. In this case outlier detection eliminates noise within the data, for instance, observations with exceptionally dissimilar information. Box plots offer a convenient visual representation of these outliers to observe the data distribution. Such plots use five statistical measures, namely: minimum value, first quartile, median (second quartile), third quartile, and maximum value. The process terms numerically distant observations within the box plot as outliers.

The most widely used outlier detection techniques are distance-based, density-based and Local Outlier Factor (LOF) algorithms. Distance-based algorithms use the distance between observations with respect to k nearest neighbors for outlier detection. An object, o, is considered to be an outlier if it does not have enough other data points as its neighbors. Density-based algorithms use data density as the criteria for outlier detection. An object, o, is considered to be an outlier if its neighborhood possesses relatively higher density. LOF, considered to be the most efficient amongst the three techniques, uses the concept of local density to detect outliers. Proposed by Breunig et al., the LOF outlier algorithm computes the average ratio of the local reachability density of an object i and it’s k Nearest Neighbors (NN) [53]. LOF can be computed using Equation (1), where, i and k are two data points.

\[
LOF_{MinPts}(i) = \frac{\sum_{k\in N_{MinPts}(i)} \frac{r_{d_{MinPts}(k)}}{r_{d_{MinPts}(i)}}}{|N_{MinPts}(i)|}
\]  

This study implements the LOF algorithm for outlier detection to processes large datasets more efficiently than the distance-based and density-based outlier detection techniques. Ali
et al. published a recent study comparing the relevance of these three techniques [54]. The resultant data contains noise-free and practical information about the building stock.

Knowledge extracted through the building stock pre-processing procedure includes geometric and non-geometric information for the entire processed building stock in a database format.

3.3. Feature Selection

Out of all the features that normally represent a building stock, only a few significantly influence building energy usage [55]. Hence, identifying the influential and optimal features is a key challenge for stakeholders. The feature selection process is one of the core concepts in the data-driven approach [28]. This paper therefore determines the key features that significantly influence building energy performance at building stock level [30]. The same features also apply in the context of the most appropriate retrofit measure to improve building energy performance. The process removes irrelevant and redundant information and only selects the set of informative features that influence building energy performance. This study implements engineering, data-driven, and hybrid approaches feature selection processes.

3.3.1. Engineering Approach

The engineering approach selects features based on existing literature and expert/survey reports published in the building energy performance domain. A human expert is the person who possesses theoretical and practical understanding of building energy performance e.g. engineer, architect, and energy modeler. Several studies identified the minimal number of crucial features that aid the energy modeling process to determine the building energy use [30]. Similarly, a limited number of studies identified the key factors that influence the energy rating of existing residential building by using EPC data [31]. These studies are either scenario specific in nature or represent a particular climate zone [27]. Implementing these features for other scenarios would lead to inaccurate results [28]. Moreover, this approach is inappropriate for datasets with a large number of features and variations.

3.3.2. Data-Driven Approach

The data-driven approach proposed in this paper determines the key features with the aid of machine learning algorithms. Identification of key features using this approach is an essential step before formulating a machine learning model. This approach is suited to large or complex datasets as these datasets often contain irrelevant and redundant features. Eliminating these features reduces the complexity of the model. Feature selection can be performed through three different methods, namely: filter, wrapper, or embedded methods [56].

Generally deployed as a pre-processing step, filter feature selection methods identify and rank features on the basis of various statistical tests. Some filter method algorithms include information gain, variance/standard deviation threshold, correlation coefficient, and chi-square tests [57]. For instance, the correlation coefficient helps filter the features that closely mirror the target feature. Similarly, a variance/standard deviation threshold filters the
features that have the most or extremely different values. The filter method is less effective than model-based methods and does not remove multi-collinearity. However, computing time is very low when compared to other methods.

Wrapper methods implement classification based learning models such as deep learning, support vector machine, and random forecast models to select the optimal features. Wrapper methods use a search algorithm that attempts to find the ‘optimal’ feature subset by selecting features based on learning model performance. The selected learning algorithm focuses on the performance of the classification-based-algorithm that has been used to predict the target value. Adopted performance indices such as precision, recall, ACCuracy (ACC), Classification Error (CE), and Root Mean Squared Error (RMSE) evaluate the effectiveness of different learning prediction models (Equations 2, 3, 4, 5, 6).

\[
\text{precision} = \frac{TP}{TP + FP} \tag{2}
\]
\[
\text{recall} = \frac{TP}{TP + FN} \tag{3}
\]
\[
\text{ACC} = \frac{TP + TN}{P + N} \tag{4}
\]
\[
\text{CE} = 1 - \frac{TP + TN}{P + N} \tag{5}
\]
\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (c_i - \bar{c}_i)^2} \tag{6}
\]

Accuracy is defined as the percentage of the correct number of predictions from all results and commonly evaluates the performance of prediction algorithms. RMSE signifies the differences between the actual and predicted outcomes. CE represents the percentage of incorrect predictions in the entire result. High recall values indicate the predicted class is correctly recognized and high precision values indicate that a case labeled as positive is indeed positive. A confusion matrix is appropriate for the visualization of algorithm performance and an overall summary of prediction results. A confusion matrix shows the summary of the total number of correct and incorrect predictions based on each output class that helps to measure the capability and scalability of a learning model.

After selection of the learning algorithm, a wrapper method identifies and selects the optimized features. This wrapper method follows an iterative process that selects a subset of features, trains them, and evaluates them based on the assigned learning algorithm performance score as shown in Figure 3. This method essentially reduces to a search problem; however, other methods are usually very computationally expensive. The process of feature selection and elimination within the subset uses heuristics with forward and backward passes. The greedy optimization algorithm is the best heuristics wrapper method to find the optimal feature subset for a large dataset. Such wrapper methods provide the optimal
subset of features whenever filter methods fail to find the optimal subset of features for any scenario. The wrapper methods also evaluate the weighted ranking of features, which indicates the influence of individual features on the target value.

Embedded methods identify the features that give the highest accuracy during the model construction process. Some examples of embedded methods include L1 (LASSO) regularization and decision trees [56].

This study employs a combination of filter and wrapper methods to determine the key features that influence building energy performance (Figure 3). These methods work particularly well with large datasets and thus, suit the scope of this study. Filter methods aid pre-processing of the involved features, which enhances the accuracy of wrapper methods [57]. This study deploys a greedy optimization coupled with the best learning algorithm to identify the optimal features [57].

The learning algorithms focus on selected features of the building stock data as identified using the filter method. The data are split into two subsets; a training set (a subset to train a model) and a test set (a subset to test the trained model). The data splitting process involves the use of one of two techniques, random data splitting and cross-validation [52]. In random data splitting, the random data are split into training and test sets using a 80/20 split respectively. Cross-validation is the most commonly used technique to achieve a balance
between the minimal bias and variance of the training model. Cross-validation divides the data into \( k \) subsets followed by the application of data splitting on each individual subset. Each iteration involves the use of a different subset. The \( k^{th} \) subset is used for testing while the other \( k - 1 \) subsets are used for training.

This study computes and predicts the energy rating values using nine different learning algorithms: 1) Deep Learning, 2) Rule Induction, 3) Neural Network, 4) Naive Bayes, 5) Decision Tree, 6) Random Forest, 7) Gradient Boosted Trees, 8) Learning Vector Quantization (LVQ) and 9) k-Nearest-Neighbours (kNN). These nine algorithms have performed excellently when used for energy forecasting \cite{6, 23}. In the context of this research, the goal is to select the best learning algorithm for the feature selection process based on the best values of the performance indices. Finally, the best learning algorithm coupled with the greedy optimization method identifies and selects the optimal features in the dataset (Figure 3).

### 3.3.3. Hybrid Approach

The hybrid approach combines engineering and data-driven approaches. The goal of this sub-process is to select the features that can be used for retrofit recommendations by using a knowledge base. The hybrid approach accomplishes feature selection in two sub-steps: 1) a combination of engineering and data-driven approaches determine the hybrid features and 2) involves selection of additional features using engineering judgement (expert opinion) after selection using data-driven approaches. These engineering features significantly influence an EPC rating and are established through expert opinions and examination of previous literature. As the data-driven approach is a black-box model, the authors note that this approach might not capture the entire spectrum of important building features. This issue arises due to the fact that the data-driven approach trains the model and removes the features solely based on model accuracy. Therefore, it is of paramount importance to manually examine features selected by the data-driven approach using engineering methods so as to retain those that are of particular relevance. In the second sub-step, a hybrid investigation facilitates selection of key features that could upgraded as part of a retrofit, these are termed ‘hybrid retrofit features’. Each selected value for a building hybrid retrofit feature is considered as a ‘single retrofit measure’. A combination of these features in the same category is termed as a ‘retrofit package’. The data-driven nature of the hybrid approach also aids the identification of ranking priority for different retrofit recommendations, as evaluated based on feature weighted rank using the wrapper method.

### 3.4. Urban Retrofit Modeling

Modeling of building retrofitting scenarios is a challenging task as a large number of factors and building features have a direct or indirect influence on building energy performance. As the feature selection procedure retains only the essential features for retrofit, this process simplifies and accelerates the process of retrofit modeling.

The creation of energy policies within the domain of urban retrofitting involves a multitude of aspects such as economic resources, environmental impact, and comfort levels. It is therefore crucial to identify the optimal combination of retrofit recommendations among
the different retrofit packages available. The developed modeling process aims to find optimum retrofit measures that not only minimize energy and CO₂ emissions but also deal with economic efficiency, enhance indoor environmental quality and long term sustainability solutions. The optimum levels of CO₂ reductions and indoor quality is associated with the target building energy rating. Similarly, Economic efficiency relates to the investment versus energy savings for long term sustainability solutions, which again is a factor of the targeted building energy rating.

Simple retrofit modeling lies in the category of shallow retrofit solutions for a building [33]. In this approach, retrofit measures often involve standard single measures to retrofit the building, such as enhancements to building fabric, lighting efficiency, and heating system efficiency [41]. Detailed retrofit modeling is an extension of simple retrofit modeling and aligns with deep retrofit modeling that follow deep retrofit guidelines to identify the most appropriate retrofit solutions [12, 38]. This procedure considers the evaluation of plausible retrofit solutions on the whole building [8]. Deep retrofitting procedures can identify solutions to target a specific energy rating.

![Diagram](image)

Figure 4: Urban scale retrofit modeling workflow for the development of a consolidated knowledge base of residential buildings

Various combinations of retrofit measures constitute a number of retrofit packages. At the building stock level, identifying the cost-optimal retrofits for a particular building involves the use of archetypes or benchmark buildings. The characteristics of the building to be retrofitted are modified and enhanced to those of the benchmark model. This work implements the real average archetype development approach to identify archetypes using the retrofit features and first classifies buildings using segmentation criteria such as dwelling type, year of construction, and heating system. Characterization of each building archetype uses aggregation based on arithmetic or geometric operations such as mean, mode, and median. The obtained aggregated values represent the characteristics of a single building archetype such as building fabric u-values and areas.

Furthermore, economic analysis of retrofit packages forms a crucial part of retrofit modeling to identify cost-effective solutions. Economic analysis evaluates the financial aspects of applying retrofit measures such as initial investment costs for the measures, annual energy savings, financial savings and financial return on the investment costs. Finally, all
the retrofit modeling results are stored in the form of a retrofit stock database (Figure 4). This information would aid policy makers and urban planners when making informed decisions and, possibly, in estimating the costs associated with large scale implementation of sustainability measures.

3.5. Building Knowledge base Development

The final step of the methodology combines and stores processed building stock data with the associated retrofit modeling measures and costs in a database. Generally, the building energy rating database is not representative of the entire building stock of a country. For instance, the Irish EPC database represents only 40% of the entire residential building stock [58]. It is necessary to analyze and evaluate current energy performance of any unrated building in order to evaluate the potential for post-retrofit energy savings. The knowledge base can be used to predict the building energy performance of the unrated building. A learning algorithm predicts existing building energy performance and recommends retrofit measures to reduce energy consumption. This algorithm can again be used to estimate the unrated building energy performance using hybrid feature values (Figure 5).

![Building knowledge-based development process with an integrated workflow to identify retrofit recommendation for a new unrated building](image)

Within the knowledge base each row refers to data for a unique building that consists of the building characteristics and retrofit stock data. The first part of the knowledge base lists all of the existing properties, characteristics and energy rating for each building based on hybrid features. The second part consists of retrofit features, targeted energy rating and total retrofit costs. The retrofit features include the target recommended values based on target energy rating (Figure 6).
4. Case Study

Use of building stock data in the energy modeling process entails various challenges, often associated with data quality and data structure. It is of paramount importance to address these quality issues before the implementation of any energy performance analysis. The formulated approach in this study addresses these quality issues through the development of a building energy performance knowledge base. This section describes the workflow of the devised methodology when applied to the EPC dataset of the Irish residential stock. The EPC data for residential buildings in Ireland are publicly available and contain an up-to-date representation of the building stock. At this point, we emphasize the fact that the implemented methodology is not case specific and can be used to develop a knowledge base for any building stock data.

Licensed EPC assessors in Ireland use the Dwelling Energy Assessment Procedure (DEAP) software for EPC estimation. This process is usually manual and hence, time intensive. Often, the estimation process requires physical measurements as inputs to DEAP, and hence, introduces human errors and data inconsistencies, such as noise. The devised methodology proposes a framework to remove these inconsistencies and furthermore, automate the process of EPC prediction using data-driven approaches. Moreover, EPC assessors also recommend plausible retrofit measures based on their expertise and DEAP calculations. The devised methodology enhances the recommendation procedure through an integrated deployment of engineering and data-driven approaches.

The test case follows the steps defined in the Section 3 and develops a knowledge base of building energy performance data for the Irish residential stock. The main steps in the database development process are now outlined.
4.1. Data Collection

Collection of the building stock data at an urban scale is quite challenging as information about any individual building is often unavailable. The data collection process involves the acquisition of raw building stock data from different sources, namely, the EPC dataset, the building census dataset, and the retrofit cost dataset (Table 2).

Table 2: Building data requirements and associated data sources for Irish case study

| Data Type                             | Case Study Data Source          | Publisher                      |
|--------------------------------------|---------------------------------|--------------------------------|
| Geometric, non-geometric             | Irish EPC database              | SEAI                           |
| and building energy                  |                                 |                                |
| performance indices                  |                                 |                                |
| Statistics                           | Irish Census database            | CSO                            |
| Retrofit Cost                        | Irish BEH and BEWH database      | SEAI                           |
| Retrofit Expert Reports              | National Housing Guideline       | Irish Department of Housing, Planning and Local Government |

The publicly available EPC data of the Irish residential stock represents the measured building stock. Published by SEAI, the EPC data contains the overall database for each residential building energy consumption data on a graded scale. An EPC for an individual house contains a building’s energy performance rating in terms of normalized primary energy consumption (kWh/(m$^2 \cdot yr$)) with the rating varying on a scale from A1 to G (Table 3). Highly efficient buildings are A1 rated and have the lowest energy consumption and CO$_2$ emissions. On the other hand, buildings with the lowest energy efficiency are G rated and hence, have the highest energy consumption and CO$_2$ emissions. The Irish EPC dataset contains approximately 850,000 residential buildings. The dataset comprises 203 building features that include building fabric, heating systems, end-use CO$_2$ emissions, delivered and primary energy consumption.

The latest statistical data (building census dataset) published by the Central Statistics Office (CSO) of Ireland contains the number of buildings in each geographical area [51]. Approximately, 1,993,672 residential buildings exist in Ireland, out of which $\approx 42\%$ have a record in the EPC dataset. The case study tests the proposed methodology using Dublin city’s EPC data, which represents $\approx 30\%$ of the EPC building stock in Ireland [58].

The retrofit cost dataset contains the finances of retrofit projects sanctioned by the Sustainable Energy Authority of Ireland (SEAI). As the costs were acquired from the recent retrofit projects such as Better Energy Homes (BEH) and Better Energy Warmer Homes (BEWH) in different counties of Ireland, these empirical costs are relevant and represent real-life scenarios. The cost data comprises of 265,182 dwelling types in BEH and 19,911 dwelling types in BEWH [49]. The authors merged this cost data from SEAI with the data from the cost optimal residential report of Ireland published by the Department of Housing, Planning and Local Government Ireland [59]. Similarly, to meet future and new
Table 3: Aggregated rating bands derived from the existing EPC rating labels for the Irish building stock to identify prediction classifiers

| Actual Rating | Energy kWh/(m² yr) | Aggregated Rating | Energy kWh/(m² yr) | Frequency | %  |
|---------------|--------------------|-------------------|--------------------|-----------|----|
| A1            | 25 or less         |                   |                    |           |    |
| A2            | 26-50              | A                 | 75 or less         | 21467     | 8.65 |
| A3            | 51-75              |                   |                    |           |    |
| B1            | 76-100             |                   |                    |           |    |
| B2            | 101-125            | B                 | 76-150             | 32099     | 12.93 |
| B3            | 126-150            |                   |                    |           |    |
| C1            | 151-175            |                   |                    |           |    |
| C2            | 176-200            | C                 | 151-225            | 80181     | 32.31 |
| C3            | 201-225            |                   |                    |           |    |
| D1            | 226-260            | D                 | 226-300            | 57355     | 23.11 |
| D2            | 261-300            |                   |                    |           |    |
| E1            | 301-340            |                   |                    |           |    |
| E2            | 341-380            | E                 | 301-380            | 29859     | 12.03 |
| F             | 381-450            | F                 | 381-450            | 13346     | 5.38 |
| G             | 451 or more        |                   | 451 or more        | 13850     | 5.58 |

building standards, the study further integrated the cost data with the data from Irish Nearly Zero Energy Building (NZEB) requirements report by SEAI [60]. According to new Irish building regulations, new buildings after 31st of December 2020 should be built according to NZEB standards (a building with a very high energy performance and a very low energy consumption). The NZEB requirements such as fabric renovation and target U-values are similar to an A-rated building. The only additional requirement is that 20% of the building energy should come from renewables, which was 10% as per the A-energy rating requirement.

4.2. Building Stock Pre-processing

The main aim of the building stock pre-processing step is to extract the Irish building characteristics and associated energy usage using data-driven methods, namely, initial statistical analysis, data pre-processing, and outlier detection. An initial statistical analysis coupled with data pre-processing of the EPC data reveals that 45 out of 203 features contain missing values. Twelve other features consist of 100% identical values (excluding the ID features). The data pre-processing step eliminates these inconsistencies; average values of features (identified using clustering of energy ratings) replace the zeroes and missing values in the dataset, while data filtering removes irrelevant data instances. Data type transformation converts all categorical and nominal values into numerical values as data-driven techniques often work with numerical values.

The density plots for roof and floor U-values reveal that the entire spectrum of U-values contain a significant number of zeroes (Figure 7). Similarly, the density plots for window,
Figure 7: Density plots of building fabric U-values (W/m²K), area (m²), and system efficiency for Dublin city consisting of more than 250,000 residential buildings with the mean value to illustrate the data distribution.
Floor, roof and door areas contain a significant number of zero values in the spectrum (Figure 7). The red bar in all the density plots represents the mean value of a feature, which has a bias towards zero due to the presence of a significant number of zeroes. Direct use of this data would lead to inaccurate results. Furthermore, the range of building heating and hot water efficiencies results also show that the efficiency field contains incorrect values; the efficiency should be between 0% to 350% because typical maximum seasonal efficiency measures for heat pumps are 350%. These findings indicate the requirement for the pre-processing step prior to use of these large datasets.

Moreover, each building falls within a construction age band, obtained through the mapping of individual years of construction into various ranges. Furthermore, a single letter (A) band represents same letter sub-band ratings (for instance, A1, A2, and A3) (Table 3). This classification would further enhance the accuracy of the learning model used in the prediction of building energy rating [61].

Outlier removal follows the data pre-processing procedure and uses the LOF algorithm. In addition, box plots graphically present the behavior of data and help identify any outliers in the data. For instance, the box plot of building element areas illustrates the effect of data pre-processing and outlier detection on individual element areas (Figure 8). The results show the presence of a significant number of outliers in the raw dataset. After the implementation of data pre-processing and outlier detection processes, the maximum value of area drops to less than 700 ($m^2$) from more than 1000 ($m^2$). Similarly, the maximum U-value undergoes a significant drop to less than 6 (W/m$^2$K) from 8 (W/m$^2$K) (Figure 9). Also, the number of outliers for building element areas are more than those for U-values.

Figure 8: Dublin EPC boxplot illustrating the building element areas ($m^2$) before and after implementation of the building stock pre-processing procedure
The processed data extracted from the EPC dataset contains only improved quantitative information for the building stock. These data are stored in a database and are also used for feature selection and retrofit modeling.

4.3. Feature Selection

The feature selection process identifies and selects only influential variables when the dataset consists of numerous features. The EPC dataset for Dublin consists of 203 features and not all of these features influence variations in the output. This study implements and compares three different feature selection approaches, namely, (1) engineering, (2) data-driven, and (3) hybrid.

The engineering approach identified 31 influential variables from the initial set of EPC data using existing studies conducted for the Irish building stock [27, 30]. Based on the investigated studies, these 31 features significantly influence the final energy consumption such as building fabric, heating systems (Table A.9). It should be noted that a common practice by the national energy agency (SEAI) when recommending a building retrofit, is to communicate a physical description of the U-value using insulation type and thickness, despite the underlying dependence between these features. Furthermore, according to a retrofit cost optimal report published by the Irish Department of Housing, Planning and Local Government, insulation thickness, insulation type and U-value of the wall need to be defined when recommending wall fabric renovation [59]. However, the selected features would differ for other case studies and available building stock data. Furthermore, the engineering features selection also depends on the user implementing the methodology to include/exclude certain features according to the modeling requirements.
The data-driven approach employs a combination of filter and wrapper methods to perform feature selection. The filter method acts as a pre-processing step and precedes the application of the wrapper method; filter method removes those features that closely mirror the target column through the use of existing correlations. The target feature in the building stock is building energy rating in terms of primary energy (kWh/(m$^2 \cdot yr$)). A correlation of less than 0.01% and more than 50% suggests that the feature has no significant influence on the building energy rating. This method also removes features that are too similar or dissimilar. A few features remain constant, with more than 90% of all values being identical. The standard deviation threshold metric eliminates these features.

The filter method identified 88 influential variables. Out of a total of 203 variables, four features have a high or low correlation with the target value. 96 other features breach the criteria as set in the filter method. A deeper investigation of the removed features indicate that floor level features such as floor fabric U-values do not fulfill the set criteria as nearly all values are identical. BER assessors, while performing surveys, often submit default values for floor level features due to the absence of accurate data. The remaining 15 features, such as, date, ID and target value have already been excluded from the analysis during the data pre-processing step.
Figure 10: Learning algorithm performance of Irish EPC dataset for the identification of the best learning algorithm

(a) Accuracy (ACC) and Classification Error (CE)

(b) Root Mean Square Error (RMSE)

The application of the wrapper method follows the filtering technique and employs a greedy optimization algorithm to determine optimal features. In this study, nine different learning algorithms such as Deep Learning, Rule Induction, Neural Network, Naive Bayes, Decision Tree, Random Forest, Gradient Boosted Trees, Learning Vector Quantization (LVQ) and K-Nearest Neighbours (KNN) compute and predict the energy rating. The processed Dublin EPC dataset tests the efficiency of these learning algorithms and uses 88 features as selected in the filter method. The data are further split into training and test sets. Data are split on the basis of a cross-validation algorithm (ten subsets of equal size). The target label defines the energy rating prediction values. The detailed application of the learning algorithms for Dublin city residential EPC data has been recently published in [55]. The selection of best learning algorithm for EPC Irish data leverages performance indices such as ACCuracy (ACC) and Classification Error (CE), and Root Mean Square Error (RMSE) (Figure 10). The result indicates that deep learning algorithms performed best when compared to other algorithms with the highest ACC of 96%, least CE of 4% and
an RMSE of 0.20. However, algorithms such as neural networks, rule induction and gradient
boosted trees also perform well with ACC values of 89%, 85% and 82% respectively.

After the selection of the best learning algorithm, the optimization algorithm identifies
the optimal features for energy rating calculations. The deep learning algorithm, coupled
with a greedy optimization algorithm, determined 41 optimal features with their weighted
ranks. That signifies the influence of each individual feature on building energy rating (Table
A.9). Some of the selected features include building fabric, heating system, and energy usage
outputs.

Table 4: Feature selection performance comparison using the deep learning algorithm for Irish EPC residential building Dublin city data

| Feature Selection Type       | ACC       | CE        | RMSE | Total Features |
|------------------------------|-----------|-----------|------|----------------|
| Data-Driven Selection        | Filter Method | 95.58% | 4.42% | 0.20           | 88             |
| Data-Driven Selection        | Wrapper Method | 82.33% | 17.67%| 0.36           | 41             |
| Engineering Selection        | 73.85% | 26.15% | 0.50 | 31             |
| Hybrid Features              | 85.85% | 14.15% | 0.34 | 56             |

A comparison of feature selection approaches based on performance indices, such as,
ACC, CE and RMSE indicate that the data-driven approach using filter methods offers the
highest accuracy of \( \approx 96\% \) although the approach requires a total of 88 features (highest
amongst all) (Table 4). On the other hand, the wrapper method offers an accuracy of \( \approx 82\% \)
and employs only 41 features. There is a significant drop in the required number of
features in the wrapper method (more than 50%) while accuracy is still high (drop of only
14%). The engineering approach achieved an accuracy of \( \approx 74\% \) using only 31 features.

Several features, such as, heating system efficiency, lighting system efficiency and re-
newable energy generation are quite influential but are not identified in the data-driven
optimization process. Renewable energy generation is also an influential feature and has a
direct relation to the building energy rating. One of the reasons for the exclusion of this
feature might be the sparse data as not many buildings have on-site renewable energy gen-
eration. As the data-driven approach relies on the supplied data to train the model, there is a
significant probability that features with sparse data will be assigned a lower weighted rank.
To cater to these limitations, the feature selection process implements the hybrid approach
that contains combined features of wrapper and engineering selection methods. The hybrid
method achieved an accuracy of \( \approx 86\% \) using only 56 features. As such, the developed
knowledge base will comprise of the 56 influential features identified. An evaluation of the
prediction model based on hybrid features indicates that individual prediction precision in
respective energy rating bands are \( \geq 71\% \) (Figure 11), which is considered to be optimal.
It is worthwhile to note that not all of these identified features can be used as retrofit measures. Hence, the feature selection process employs a hybrid retrofit feature selection approach that comprises sixteen influential features for building retrofitting. The first set of nine features are common to both data-driven and engineering selection methods. The process extracts the second set of seven features from existing retrofit studies that represent frequently implemented retrofit recommendations. These sixteen features can be categorized on the basis of retrofit type and retrofit priority (Table 5). The retrofit type consists of four main categories of retrofits, namely, building fabric, heating system, lighting and renewables. The retrofit priority comprises of labels from 1 to 9 formulated on the basis of a weighted rank. Labels 1-9 represent the priority of the first set of nine features. The second set of seven features do not have any relative assigned priority (Table 5). The retrofit features represent specific variables of shallow retrofit implementation, which have an associated priority. Furthermore, it is worthwhile to mention that the data-driven modeling process omits certain crucial variables such as ventilation and low energy lighting from the list of recommended features. Therefore, these features are added to the feature list without any associated priority. Retrofit priority is determined using the wrapper method, which evaluates the weighted ranking of retrofit features indicating the influence of individual features on the target energy rating.
| Retrofit Features                  | Description                  | Retrofit Type    | Retrofit Priority |
|-----------------------------------|------------------------------|------------------|-------------------|
| Main Water Heating Fuel           | Hot water heating fuel       | Heating System   | 1                 |
| Structure Type                    | Structure type              | Fabric           | 2                 |
| Insulation Type                   | Insulation type             | Fabric           | 3                 |
| Thermal Mass Category             | Thermal mass                | Fabric           | 4                 |
| U-Value Wall                      | U-value wall                | Fabric           | 5                 |
| U-Value Roof                      | U-value roof                | Fabric           | 6                 |
| U-Value Floor                     | U-value floor               | Fabric           | 7                 |
| U-Value Window                    | U-value window              | Fabric           | 8                 |
| Heating Efficiency                | Heating system efficiency   | Heating System   | 9                 |
| Insulation Thickness              | Insulation thickness        | Fabric           | -                 |
| Main Space Heating Fuel           | Heating system fuel         | Heating System   | -                 |
| Ventilation Method                | Ventilation method          | Heating System   | -                 |
| Hot water Efficiency              | Hot water system efficiency | Heating System   | -                 |
| Low Energy Lighting Percent       | Low energy lighting         | Lighting         | -                 |
| Solar Hot Water Heating           | Solar hot water heating     | Renewables       | -                 |
| Solar PV                          | Solar photovoltaic panels   | Renewables       | -                 |

The process of feature selection for the Irish EPC residential building data experiences a gradual decline in the number of features (Figure 12): the initial step of data collection consists of 203 features, which is reduced to a mere 16 features in the last step of hybrid selection. The number of features drop down to 188 after the building stock pre-processing procedure (step 2). Implementation of the filter method (step 3) further reduces the number of relevant features to 88. Wrapper methods identify a set of 41 crucial features (step 4). The number of features identified using the engineering selection (step 5) and hybrid selection techniques (step 6) are 31 and 56 respectively. Step 7 identifies features that can be recommended as retrofit measures, which reduces the number to 16.
4.4. Urban Retrofit Modeling

Urban retrofit modeling follows the feature selection process, which identifies influential features to be used as potential retrofit recommendations. The retrofit modeling process classifies all the target features into different packages (P1-P9) to identify the associated retrofit costs based on the retrofit cost optimal report (Table 6). There are four individual measure types classified as shallow retrofit recommendations; different combinations of these measures represent measures of deep retrofitting. For instance, entire heating and hot water systems will be upgraded in the heating system type retrofit measure (labeled as Package P1). The building fabric renovation comprises of two packages, namely, P2 and P3. The P2 package involves the renovation of walls, roof, and floor while the P3 package involves retrofit recommendations, such as, window renovation. Similarly, packages P4 and P5 comprise of renewable system installation and lighting system retrofits respectively. A combination of different packages forms four more packages (Packages P6 to P9). For instance, package P9 covers the entire building renovation (heating system, building fabric, renewables and lighting).

The identified packages contain one or more of 16 influential features, which are the potential retrofit measures. The identified influential features consist of thermal as well as non-thermal variables. For instance, window U-values and insulation thickness are parts of the thermal inputs, while non-thermal inputs comprise for example: low energy lighting percent and hot water efficiency. It is worthwhile to mention that inputs influencing the EUI mostly constitute thermal variables irrespective of the approach employed to identify these variables. Furthermore, the identified retrofit packages are dominated by renovation...
strategies focusing on thermal variables. For instance, packages P2, P3, P6, P7, P8 and P9 consist of fabric renovation strategies which involve variables with a direct influence on the thermal behavior. Moreover, fabric renovation also translates to a substantial amount of investment when considering individual packages P1 to P5.

Table 6: Detailed description of retrofit packages identified in the retrofit modeling process to suggest retrofit recommendations

| Retrofit Packages | Retrofit Measures Type | Retrofit Recommendations |
|-------------------|------------------------|--------------------------|
| P1                | Heating System         | Heating and hot water with ventilation upgrade |
| P2                | Fabric                 | Wall, roof and floor renovation |
| P3                | Fabric                 | Window renovation |
| P4                | Renewables             | Solar PV and hot water installation |
| P5                | Lighting               | Lighting upgrade |
| P6                | Heating System, Fabric | P1 and P2 |
| P7                | Heating System, Fabric | P1 and P3 |
| P8                | Heating System, Fabric, Renewables | P1, P2 and P4 |
| P9                | Heating System, Fabric, Lighting, Renewables | P1, P3, P4 and P5 |

Determination of existing performance features for each building is quite challenging at an urban scale. This study implements the buildings archetypes approach to obtain values of various building features. The archetype development process classifies and groups different buildings on the basis of their respective energy ratings using simple aggregation (Table 7). An archetype represents buildings with a similar energy rating. For instance, buildings in the A rated archetype have a window U-value of 1.26 W/m².K, floor U-value of 0.14 W/m².K and so on (column 3 of Table 7).
Table 7: Characteristic values of building archetypes' features based on energy ratings from Irish EPC Dublin city data; individual feature values used as a reference to identify the target values for upgrading any particular building from a lower energy rating to a higher one

| Features                  | Unit | A   | B   | C   | D   | E   | F   | G   |
|---------------------------|------|-----|-----|-----|-----|-----|-----|-----|
| U-value window            | W/m²·K | 1.26 | 2.10 | 2.70 | 2.80 | 3.02 | 3.19 | 3.70 |
| U-value floor             | W/m²·K | 0.14 | 0.34 | 0.48 | 0.55 | 0.62 | 0.68 | 0.73 |
| U-value roof              | W/m²·K | 0.14 | 0.22 | 0.31 | 0.40 | 0.57 | 0.99 | 1.64 |
| U-value wall              | W/m²·K | 0.17 | 0.36 | 0.55 | 0.60 | 1.49 | 1.86 | 1.97 |
| Insulation thickness      | mm   | 53  | 38  | 35  | 35  | 35  | 38  | 36  |
| Structure type            |      | MA  | MA  | MA  | MA  | MA  | MA  | MA  |
| Insulation type           |      | FI  | FI  | FI  | FI  | FI  | LJ  | LJ  |
| Thermal mass              |      | H   | M   | M   | M   | M   | M   | M   |
| Heating system fuel       |      | Gas | Gas | Gas | Gas | Gas | Gas | Elec |
| Hot Water heating fuel    |      | Gas | Gas | Gas | Gas | Gas | Gas | Elec |
| Ventilation method        |      | Bal | Nat | Nat | Nat | Nat | Nat | Nat |
| Hot water efficiency      | %    | 91.10 | 90.10 | 89.90 | 85.00 | 79.70 | 78.10 | 84.00 |
| Heating system efficiency | %    | 91.10 | 90.20 | 89.70 | 84.86 | 79.60 | 78.00 | 78.80 |
| Low energy lighting       | %    | 100.00 | 66.18 | 50.19 | 44.08 | 41.25 | 38.74 | 41.14 |
| Solar PV                  | boolean | Yes | No  | No  | No  | No  | No  | No  |
| Solar hot water heating   | fraction | 42.37 | 28.85 | 18.98 | 7.82 | 2.36 | 0.45 | 0.17 |
| EUI                       | kWh/m²·yr | 53  | 129 | 189 | 259 | 336 | 414 | 544 |
| Total Buildings           |      | 21467 | 32099 | 80181 | 57355 | 29859 | 13850 | 13346 |

MA, Masonry; FI, Factory Insulated; LG, Loose Jacket; H, High; M, Medium; Elec, Electricity; Bal, Balance mechanical ventilation; Nat, Natural ventilation; EUI, Energy Use Intensity

These characteristic values also represent a reference to infer the target values for upgrading any particular building from a lower energy rating to a higher one (Table 7). For instance, to upgrade a G-rated building to an A-rated building, it would be necessary to decrease the U-value of the wall from 1.97 to 0.17. Similarly, it becomes relatively straightforward to identify the target values of different building features for renovation.

An analysis of the building archetypes indicates that some features have similar values for different energy ratings, such as, building structure type, ventilation method and thermal mass. For instance, gas is the main source of fuel for space heating and hot water systems. Similarly, most of the buildings use natural ventilation as the main technique for ventilation. Alongside, the most commonly used building material has medium thermal mass with
Building in the lower rating bands (C, D, E, F, and G) have an average lighting efficiency of \( \leq 50\% \). As lighting efficiency is one of the influential factors in energy rating calculation, low energy lighting systems would be a potential retrofit recommendation to enhance the existing rating. Furthermore, most of the A-rated buildings have an on-site solar PV system indicating that renewables have a significant impact on the building energy performance, even when the buildings are well-insulated. Approximately, 42% and 28% of the A-rated and B-rated buildings have an on-site solar water heating system. Although this study employs only energy rating classification for archetypes development, energy rating and dwelling type classification criteria can also be used to obtain a detailed archetypal classification (Table B.10).

This study also implements the Irish NZEB standards in the retrofit analysis case study. The NZEB requirements are quite similar to an A1 rated building for fabric renovation. Heat pumps are the ideal heating system for NZEB buildings. Also, 20% of the building energy should come from renewables.

The retrofit modeling process concludes with an economic analysis of the various retrofit possibilities at the urban scale. The economic analysis evaluates the developed retrofit packages (Table 6) in terms of their financial viability using the identified building archetype characteristics (Table 7). As expected, the costs experience a steep increase when upgrading B, C, D, E, F, and G rated buildings to NZEB standards (Table 8). It is worthwhile to mention that the total cost (column II in Table 8) corresponds to costs associated with the implementation of P9 retrofit package. The shift in energy rating (column I in Table 8) is only possible when all existing building features include the target features of P9 package. Moreover, renewables are only considered in the analysis when the target is an A-rated or NZEB building. The P1 package for heating or hot water system upgrade considers the boiler replacement on the basis of the target energy efficiency. However, as a heat pump is the optimal choice for NZEB and A-rated targets, the implementation of a heat pump system adds an extra investment of \( \approx €9000-13000 \). Furthermore, building fabric renovation (P2 package) also entails a significant amount of investment (column IV in Table 8). An A-rated building has similar characteristics as the NZEB building. Hence, there is no associated cost of investment for upgrading A-rated buildings to NZEB. The only additional investment is the installation of renewable generation sources (package P4). Furthermore, A-rated and NZEB buildings employ similar balanced ventilation systems. Hence, the investment for the P1 package (upgrading A1 to NZEB) has a zero value.
Table 8: Approximate average cost (€) for upgrading lower rated buildings to higher rated ones considering P1 to P9 retrofit packages

| Energy Rating | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
|---------------|----|----|----|----|----|----|----|----|----|
| A to NZEB     | 0  | 0  | 0  | 5000 | 0  | 0  | 0  | 5000 | 5000 |
| B to NZEB     | 9000 | 14978 | 6086 | 10000 | 150 | 23978 | 15086 | 33978 | 40214 |
| C to NZEB     | 11000 | 24338 | 6086 | 10000 | 150 | 35338 | 17086 | 45338 | 51574 |
| D to NZEB     | 11000 | 24338 | 6086 | 10000 | 150 | 35338 | 17086 | 45338 | 51574 |
| E to NZEB     | 12000 | 26211 | 6086 | 10000 | 150 | 38211 | 18086 | 48211 | 54447 |
| F to NZEB     | 13000 | 28083 | 6086 | 10000 | 150 | 41083 | 19086 | 51083 | 57319 |
| G to NZEB     | 13000 | 29955 | 6086 | 10000 | 150 | 42955 | 19086 | 52955 | 59191 |
| B to A        | 9000 | 14978 | 6086 | 10000 | 150 | 23978 | 15086 | 33978 | 40214 |
| C to A        | 11000 | 24338 | 6086 | 10000 | 150 | 35338 | 17086 | 45338 | 51574 |
| C to B        | 7000 | 7489 | 5326 | 0    | 80  | 14489 | 12326 | 14889 | 19895 |
| D to A        | 11000 | 24338 | 6086 | 10000 | 150 | 35338 | 17086 | 45338 | 51574 |
| D to B        | 7000 | 7489 | 5326 | 0    | 60  | 14489 | 12326 | 14889 | 19875 |
| D to C        | 6000 | 3744 | 4565 | 0    | 50  | 9744  | 10565 | 9744  | 14399 |
| E to A        | 12000 | 26211 | 6086 | 10000 | 150 | 38211 | 18086 | 48211 | 54447 |
| E to B        | 7000 | 14978 | 5326 | 0    | 80  | 21978 | 12326 | 21978 | 27384 |
| E to C        | 6000 | 9361 | 4565 | 0    | 0   | 15361 | 10565 | 15361 | 19926 |
| E to D        | 5000 | 9361 | 3804 | 0    | 0   | 14361 | 8804  | 14361 | 18165 |
| F to A        | 13000 | 28083 | 6086 | 10000 | 150 | 41083 | 19086 | 51083 | 57319 |
| F to B        | 7000 | 16850 | 5326 | 0    | 80  | 23850 | 12326 | 23850 | 29256 |
| F to C        | 6000 | 11233 | 4565 | 0    | 50  | 17233 | 10565 | 17233 | 21848 |
| F to D        | 5000 | 7489 | 3804 | 0    | 0   | 12489 | 8804  | 12489 | 16293 |
| F to E        | 0    | 5617 | 3043 | 0    | 0   | 5617  | 3043  | 5617  | 8660 |
| G to A        | 13000 | 29955 | 6086 | 10000 | 150 | 42955 | 19086 | 52955 | 59191 |
| G to B        | 7000 | 16850 | 5326 | 0    | 80  | 23850 | 12326 | 23850 | 29256 |
| G to C        | 6000 | 11233 | 4565 | 0    | 50  | 17233 | 10565 | 17233 | 21848 |
| G to D        | 5000 | 7489 | 3804 | 0    | 0   | 12489 | 8804  | 12489 | 16293 |
| G to E        | 0    | 5617 | 3043 | 0    | 0   | 5617  | 3043  | 5617  | 8660 |
| G to F        | 0    | 5617 | 2282 | 0    | 0   | 5617  | 2282  | 5617  | 7899 |

P9, Total cost required for target rating; P1, Cost of heat pump installation to target NZEB or A-rated building and boiler replacement cost for other ratings; P2, Renovation cost based on building fabric area; P3, Renovation cost based on window area; P4, Whole renewable system installation cost

4.5. Building Knowledge base Development

The building knowledge base development process merges the processed Irish residential building data, retrofit modeling results, and associated economic analysis into one knowledge base. Each row in the knowledge base refers to a unique building and identifies two sources of information. The first source lists the existing characteristics of each building, which includes a total of 56 features identified on the basis of a hybrid selection process (data-driven and engineering selection approaches) as discussed in Section 4.3. The second source of information includes the retrofit features and retrofit costs. The retrofit features set the target references to achieve different target energy ratings (Table 7). The retrofit cost associates these features with the costs for upgrading existing buildings to the target values.
This knowledge base could find use in various applications in the urban energy building simulation domain. For instance, decision makers could use the knowledge base to predict the current building energy performance of any building and hence, evaluate different retrofit alternatives. The deep learning algorithm used for energy rating prediction attains a high accuracy with minimum input features. As illustrated in the hybrid feature selection process, only 56 features can predict the building energy performance as opposed to the 203 features used by SEAI for EPC calculation.

The knowledge base can also be used to suggest retrofit recommendation. As the knowledge base consists of retrofit features, these can be used to achieve certain target ratings. For instance, the knowledge base will contain the target values of features to enhance a C-rated building to an A-rated one. Alongside, the knowledge base will also suggest the investment required for the upgrade.

The real-life application of the knowledge-base could be further extended to optimize large-scale retrofit implementation, particularly at an urban scale. This application can be illustrated using the same case study to estimate the required retrofit investment costs (Figure 13a), CO₂ reductions (Figure 13b) and EUI savings (Figure 13c) using the target retrofit features of NZEB, A, and B ratings. For instance, retrofit investment costs give an indication of the investment costs when upgrading A-rated buildings to NZEB standards, B-rated buildings to A-rated and NZEB standards, C-rated buildings to A-rated, B-rated and NZEB standards and so on at the Dublin city scale (Figure 13a). All of these calculations correspond to retrofit features defined in the P9 package when identifying the cost estimates. To identify the number of buildings that fall into a particular rating band, this process superimposes individual rating distribution weights of the EPC dataset on the CSO dataset. The results indicate that upgrades to buildings in lower rating bands to NZEB, A-rated, and B-rated buildings require an investment of approximately €9977 million, €9885 million and €3708 million respectively. To upgrade the buildings in the upper bands (A and B rated buildings to NZEBs), the investment cost is relatively lower than to upgrade the buildings in the lower bands (E, F and G-rated buildings). For instance, upgrading C and D rated buildings requires higher investments because buildings in C and D rated bands have a higher distribution weight than any other rating band in Dublin city. These estimates would help the urban planners in targeting renovations in areas of special interest. Furthermore, the policy makers would be able to keep a track of the builder sector in terms of energy efficiency and carbon emissions.
5. Discussion

Although the proposed methodology can be applied generally, one of the significant challenges with this approach is the availability of building stock data. Furthermore, the scope and quality of the available data also has a huge impact on the generated results as
the approach has a high dependency on data. These limitations are expressed in further detail as follows.

- **EPC Coverage Bias:** The EPC data should be representative of the entire building stock to propose inferences about similar property types in a given area. Often, rated dwellings may differ from unrated dwellings in a systematic way that correlates with energy use/energy efficiency. For example, in the Irish context, it is mandatory for a dwelling to have a valid EPC when it is available to rent or buy. It is likely that dwellings with a poor energy rating are less likely to be sold or rented, and hence, these buildings are less likely to have an EPC. This creates a bias whereby the housing stock appears more energy efficient while the analysis only considers the dwellings with an EPC. This also suggests that areas with a low overall EPC coverage, but with a sizeable number of new homes, will appear to be more energy efficient as new homes are built to higher energy efficiency standards and need to have a EPC at the time of purchase. Policymakers, using the underlying data or the mapping outputs, should be mindful of the representativeness of the data in a given area.

- **The Gap Between Estimated Energy Use and Actual Energy Use:** Another limitation of EPC data is that there is often a large gap between estimated (also referred to as calculated energy use) and actual energy use. The EPC of a building is based on the estimated energy use that a building is likely to use given its physical parameters. A physical model relies on common assumptions about factors such as occupancy to evaluate the energy consumption of a dwelling. Estimated energy use, therefore, ignores occupant behaviour, which often results in over estimation of energy consumption (pre-bound effect). The estimation process is unable to incorporate the demand patterns of individual households. These differences can make it difficult to predict the actual energy usage in a given area based on estimated energy use data as proposed in this study.

- **Building Archetype Development Approach Bias:** As the archetype development process employs a simple aggregation technique, this might lead to biased retrofit stock modeling results. The employed aggregation approach might eliminate crucial building characteristic values. For instance, a heat pump delivers high efficiency when compared to a boiler. However, the aggregation results indicate that boilers are the most frequently used heating systems for A-rated buildings. This is due to the fact that only a limited number of A-rated buildings use heat pumps for space heating.

Furthermore, selection of the segmentation criteria for archetype development might also affect the building classification. For instance, when developing archetypes on the basis of construction year, current retrofit status should also be taken into account. All the retrofitted buildings should be classified through an enhanced construction band. This case study dataset does not contain any information regarding the current retrofit status of the building. Therefore, any classification based on year of construction would be misleading for this particular case study. However, for any general case, year of construction could be used to predict the energy rating of the building provided
that the retrofitted buildings are treated as outliers and represented using a classifier. Moreover, it would be quite obvious that the old non-retrofitted buildings would fall in the lower band of rating classification and the new buildings would fall in the upper band of rating classification as new buildings tend to follow improved standards of construction practice (Table B.11).

Hence, it is fairly crucial to analyze the input data for any potential bias or assumptions so that any decisions made using the output represent similar bias and assumptions.

6. Conclusions and Future Work

One of the leading approaches for achieving a sustainable environment is retrofitting of existing buildings. As retrofit measure identification requires a baseline for existing building performance, a significant challenge for any stakeholder is to extract useful knowledge from existing available data for urban-scale retrofit modeling. The main aim of this study is to develop a generalized methodology to optimize urban scale energy retrofit decisions for residential buildings using data-driven approaches. Stakeholders can estimate the required retrofit investment costs, emission reductions and energy savings using the target retrofit features of energy-efficiency ratings. Given sufficient building coverage, this approach may help urban planners assess energy consumption in a given area. This would eventually aid the urban planners and policy makers in the identification of optimal locations for large infrastructure projects such as district heating.

The methodology devised in this study could aid in the development of a consolidated knowledge-base as demonstrated for the considered case study. Such a consolidated knowledge-base could contain information regarding available Energy Performance Certificates of existing buildings, census reports, retrofit case studies, influential features, retrofit recommendations and associated implementation costs. It is worthwhile to mention that the majority of energy modeling gaps at the urban level deal with data related issues with the data being limited or almost non-existent. Development of a consolidated knowledge-base would address these issues and therefore, substantially enhance the accuracy of urban scale energy modeling. Furthermore, it is of paramount importance that the knowledge-base comprises only crucial information with variables that highly influence energy consumption. A feature selection process as demonstrated in the case study helps in eliminating any irrelevant or redundant information. This process further adds additional value to the knowledge-base in the form of weighted ranks representing the priorities of retrofit implementation. As the knowledge-base comprises both building and retrofit stock data, such a database would help stakeholders to evaluate the energy performance of residential buildings and find feasible retrofit solutions by leveraging scarce available resources at an urban scale. The results further indicate that implementation of the data-driven approach enhances the quality of existing data related to buildings and is instrumental in the extraction of key features from complex data.

Generally, the building stock information for any country exists as an energy performance certificate database, which aids in the calculation of existing building energy performance
baselines for retrofit analysis. For instance, the energy performance certificate database of Dublin represents 30% of the Irish residential building stock, which consists of 203 building features. One significant challenge for stakeholders is the identification of retrofit strategies with limited sources for the entire residential stock of Dublin. The proposed methodology extracts the knowledge from available resources to identify the existing building energy performance and formulate retrofit solutions. Results signify the importance of data-driven retrofit modeling as the feature selection process reduces the number of features in Dublin’s building stock database from 203 to 56 with a building rating prediction accuracy of 86%. Amongst the 56 features, 16 identify as recommended retrofit measures for each energy-efficiency rating. Furthermore, the process evaluates the implementation costs associated with each retrofit measure. As such, Dublin city planners and policymakers (Sustainable Energy Authority Of Ireland) can estimate the required retrofit investment costs, emission reductions and energy savings using the target retrofit features of energy-efficiency ratings.

Future work will investigate whether further improvements can be achieved by integrating planning data with geographical mapping. The research could be further expanded to include commercial buildings. Furthermore, there is a need for a social science (such as occupancy and demand patterns) based research to investigate how buildings with the same Energy Performance Certificate differ statistically in terms of their measured consumption. Also, it would be interesting to investigate why some households that could afford better thermal comfort standards choose to use so little heating energy in non-retrofitted homes.

7. Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under the SFI Strategic Partnership Programme Grant number SFI/15/SPP/E3125. We acknowledge the Sustainable Energy Authority of Ireland (SEAI) for access to anonymised Building Energy Rating (BER), Better Energy Homes (BEH) and Better Energy Warmer Home (BEWH) datasets. 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 SFI and SEAI. This work emerged from the IBPSA Project 1, an international project conducted under the umbrella of the International Building Performance Simulation Association (IBPSA).
### Appendix A. Irish EPC Data Features

Table A.9: Selected features based on data-driven and engineering selection methods

| Features                          | Data-Driven Selection | Engineering Selection |
|----------------------------------|-----------------------|-----------------------|
| Year of Construction             | 1                     | Door Area             |
| Predominant Roof Type            | 2                     | Dwelling Type         |
| Second Wall Is Semi Exposed      | 3                     | Floor Area            |
| Main Water Heating Fuel          | 4                     | Heating Efficiency    |
| Structure Type                   | 5                     | Heating Supply Heat Fraction |
| Electricity Immersion In Summer  | 6                     | Heating Supply System Efficiency |
| Insulation Type                  | 7                     | Insulation Thickness  |
| Primary Circuit Loss             | 8                     | Insulation Type       |
| Thermal Mass Category            | 9                     | Living Area Percent   |
| Ground Floor Area(sq m)         | 10                    | Main Space Heating Fuel|
| U-Value Wall                     | 11                    | Main Hot Water Fuel   |
| U-Value Roof                     | 12                    | No Storeys            |
| U-Value Floor                    | 13                    | Low Energy Lighting Percent |
| U-Value Window                   | 14                    | Roof Area             |
| Door Area                        | 15                    | Room In Roof Area     |
| Heating Main System Efficiency   | 16                    | Structure Type        |
| Maximum Permissible CO2 Emission Rate | 17                | Supply Heating Fuel   |
| Heating Supply Heat Fraction     | 18                    | Supply Hot Water Fuel |
| Hot Water Efficiency Adj Factor  | 19                    | Thermal Mass Category |
| Delivered Lighting Energy        | 20                    | U-Value Door          |
| Delivered Energy Main Space      | 21                    | U-Value Floor         |
| Primary Energy Lighting          | 22                    | U-Value Roof          |
| Primary Energy Pumps Fans        | 23                    | U-Value Wall          |
| Primary Energy Main Water        | 24                    | U-Value Window        |
| Primary Energy Main Space        | 25                    | Ventilation Method    |
| CO2 Main Space                   | 26                    | Wall Area             |
| Ground Floor Height              | 27                    | Hot water Efficiency  |
| Total Delivered Energy           | 28                    | Window Area           |
| Living Area Percent              | 29                    | Year of Construction  |
| CO2 Secondary Space              | 30                    | Solar Hot Water Heating|
| CO2 Supplementary Water          | 31                    | Solar PV              |
| Primary Energy Secondary Space   | 32                    |                      |
| First Wall Area                  | 33                    |                      |
| First Wall U-Value               | 34                    |                      |
| No Storeys                       | 35                    |                      |
| Supply Heating Fuel              | 36                    |                      |
| Suppl Hot Water Fuel             | 37                    |                      |
| No Of Chimneys                   | 38                    |                      |
| No Of Open Flues                 | 39                    |                      |
| No Of Fans And Vents             | 40                    |                      |
| Percentage Draught Stripped      | 41                    |                      |

39
### Appendix B. Building Archetypes Examples

Table B.10: Building Archetypes U-values characteristics in W/m².K based on Energy Rating and Dwelling Types

| Energy Rating | A   | B   | C   | D   | E   | F   | G   |
|---------------|-----|-----|-----|-----|-----|-----|-----|
| **Semi-detached house** |     |     |     |     |     |     |     |
| Window        | 1.24| 2.00| 2.70| 2.80| 3.10| 3.36| 3.91|
| Floor         | 0.14| 0.36| 0.46| 0.57| 0.65| 0.66| 0.73|
| Roof          | 0.13| 0.20| 0.31| 0.40| 0.49| 0.74| 1.54|
| Wall          | 0.17| 0.29| 0.55| 0.60| 1.78| 2.01| 2.08|
| **Detached house** |     |     |     |     |     |     |     |
| Window        | 1.24| 2.04| 2.70| 2.82| 3.07| 3.28| 3.88|
| Floor         | 0.14| 0.35| 0.47| 0.59| 0.65| 0.70| 0.73|
| Roof          | 0.14| 0.21| 0.29| 0.40| 0.49| 0.72| 1.62|
| Wall          | 0.17| 0.30| 0.55| 0.60| 1.64| 1.91| 2.04|
| **House**     |     |     |     |     |     |     |     |
| Window        | 1.54| 1.78| 2.70| 2.80| 3.10| 3.10| 4.24|
| Floor         | 0.16| 0.25| 0.41| 0.56| 0.59| 0.65| 0.71|
| Roof          | 0.15| 0.17| 0.26| 0.40| 0.40| 0.87| 1.63|
| Wall          | 0.20| 0.27| 0.54| 0.60| 1.64| 1.78| 1.98|
| **Apartment** |     |     |     |     |     |     |     |
| Window        | 1.29| 2.00| 2.70| 2.80| 2.80| 2.86| 3.10|
| Floor         | 0.15| 0.26| 0.36| 0.41| 0.48| 0.59| 0.73|
| Roof          | 0.18| 0.24| 0.34| 0.49| 0.71| 1.03| 1.56|
| Wall          | 0.17| 0.36| 0.53| 0.55| 0.60| 0.87| 1.57|
| **Terrace house** |     |     |     |     |     |     |     |
| Window        | 1.28| 2.20| 2.70| 2.87| 3.17| 3.61| 4.36|
| Floor         | 0.13| 0.31| 0.46| 0.49| 0.57| 0.61| 0.61|
| Roof          | 0.13| 0.18| 0.25| 0.40| 0.67| 1.27| 2.30|
| Wall          | 0.16| 0.35| 0.53| 1.00| 1.78| 2.06| 2.10|
Table B.11: Semi-detached house archetype window U-value characteristics in W/m².K based on energy rating and year of construction

| Year of Construction | A    | B    | C    | D    | E    | F    | G    |
|----------------------|------|------|------|------|------|------|------|
| 1899 or Less         | -    | 2.86 | 3.51 | 3.57 | 3.86 | 4.01 |
| 1900-1929            | -    | 1.89 | 2.46 | 2.89 | 3.46 | 3.83 | 3.94 |
| 1930-1949            | -    | 1.85 | 2.46 | 2.83 | 3.29 | 3.68 | 4.09 |
| 1950-1966            | -    | 1.86 | 2.50 | 2.87 | 3.33 | 3.69 | 4.11 |
| 1967-1977            | -    | 2.00 | 2.57 | 2.90 | 3.15 | 3.52 | 3.84 |
| 1978-1982            | -    | 2.09 | 2.68 | 3.04 | 3.27 | 3.76 | 4.18 |
| 1983-1993            | -    | 2.21 | 2.70 | 3.03 | 3.48 | 3.56 | 3.13 |
| 1994-2004            | -    | 2.20 | 2.73 | 2.87 | 2.94 | 2.93 |    |
| 2005-2010            | -    | 1.81 | 2.30 | 2.43 |    |    |    |
| 2011 or above        | 1.21 | 1.47 |    |    |    |    |    |

A brief analysis of the window U-values suggests a gradual decline until the year band 1994-2004. Building regulations were introduced in Ireland after 2002 and hence, the U-values experience a steep decline for all energy ratings between 2005 and 2010. Furthermore, buildings constructed in after 2011 should be built to a minimum of B rating according to the new construction standards. Therefore, this year band only consists of A and B-rated buildings.

References

[1] Energy for Europe by European Commission, [https://ec.europa.eu/energy/en](https://ec.europa.eu/energy/en), [Online; accessed 2-Mar-2020].

[2] T. Abergel, B. Dean, J. Dulac, Towards a zero-emission, efficient, and resilient buildings and construction sector: Global status report 2017, UN Environment and International Energy Agency: Paris, France.

[3] E. Recast, Directive 2010/31/eu of the European parliament and of the council of 19 may 2010 on the energy performance of buildings (recast), Official Journal of the European Union 18 (06) (2010) 2010.

[4] EU, Directive (EU) 2018/844 of the European Parliament and of the Council of 30 may 2018 amending Directive 2010/31/eu on the energy performance of buildings and Directive 2012/27/eu on energy efficiency, Official Journal of the European Union 61.

[5] J. Kneifel, Life-cycle carbon and cost analysis of energy efficiency measures in new commercial buildings, Energy and Buildings 42 (3) (2010) 333–340.

[6] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047.

[7] É. Mata, A. Sasic Kalagasidis, F. Johnsson, Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK, Building and Environment 81 (2014) 270–282.

[8] A. G. Charalambides, C. N. Maxoulis, O. Kyriacou, E. Blakeley, L. S. Frances, The impact of energy performance certificates on building deep energy renovation targets, International Journal of Sustainable Energy 38 (1) (2019) 1–12.

[9] U. EIA, Commercial buildings energy consumption survey, United Sates Department of Energy, Ed., ed.
[10] C. S. Monteiro, C. Costa, A. Pina, M. Y. Santos, P. Ferrão, An urban building database (ubd) supporting a smart city information system, Energy and Buildings 158 (2018) 244–260.

[11] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrkani, O. Kordas, Energy performance certificates new opportunities for data-enabled urban energy policy instruments?, Energy policy 127 (2019) 486–499.

[12] N. Strazza, P. Sdrigotti, C. A. Stival, R. Berto, Evaluating deep retrofit strategies for buildings in urban waterfronts, in: Sustainable Building for a Cleaner Environment, Springer, 2019, pp. 391–403.

[13] R. Ruparethna, K. Hewage, R. Sadiq, Improving the energy efficiency of the existing building stock: A critical review of commercial and institutional buildings, Renewable and sustainable energy reviews 53 (2016) 1032–1045.

[14] Building Performance Database by Department of Energy, https://bpd.lbl.gov/ [Online; accessed 2-Mar-2020].

[15] Censushub - European Statistical System (ESS), https://ec.europa.eu/CensusHub2 [Online; accessed 2-Mar-2020].

[16] Data hub for energy performance of buildings, buildings performance institute europe (BPIE), http://bpie.eu/ [Online; accessed 2-Mar-2020] (2017).

[17] Assessment and improvement of the EPBD impact (for new buildings and building renovation) (ASIEPI), Intelligent Energy Europe- European Commission, https://ec.europa.eu/energy/intelligent/projects/en/projects/asiepi [Online; accessed 2-Mar-2020].

[18] Energy Performance Certificates across europe from design to implementation, BPIE, http://www.bup.eu [Online; accessed 2-Mar-2020] (2010).

[19] M. Collins, J. Curtis, Bunching of residential building energy performance certificates at threshold values, Applied Energy 211 (2018) 662–676.

[20] J. von Platten, C. Holmberg, M. Mangold, T. Johansson, K. Mjörnell, The renewing of energy performance certificatesreaching comparability between decade-apart energy records, Applied Energy 255 (2019) 113902.

[21] J. Keirstead, M. Jennings, A. Sivakumar, A review of urban energy system models: Approaches, challenges and opportunities, Renewable and Sustainable Energy Reviews 16 (6) (2012) 3847–3866.

[22] C. F. Reinhart, C. C. Davila, Urban building energy modeling–a review of a nascent field, Building and Environment 97 (2016) 196–202.

[23] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205.

[24] A. Nutkiewicz, Z. Yang, R. K. Jain, Data-driven urban energy simulation (due-s): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow, Applied energy 225 (2018) 1176–1189.

[25] N. Abbaspabadi, M. Ashayeri, R. Azari, B. Stephens, M. Heidarinejad, An integrated data-driven framework for urban energy use modeling (neum), Applied energy 253 (2019) 113550.

[26] 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 and Buildings 165 (2018) 301–320.

[27] A. A. Famuyibo, A. Duffy, P. Strachan, Developing archetypes for domestic dwellingsan irish case study, Energy and Buildings 50 (2012) 150–157.

[28] C. Fan, Y. Sun, Y. Zhao, M. Song, J. Wang, Deep learning-based feature engineering methods for improved building energy prediction, Applied energy 240 (2019) 35–45.

[29] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustainable Cities and Society 39 (2018) 508–518.

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

[31] A. Stone, D. Shipworth, P. Biddulph, T. Oreszczyn, Key factors determining the energy rating of existing english houses, Building Research & Information 42 (6) (2014) 725–738.

[32] L. Zhang, J. Wen, A systematic feature selection procedure for short-term data-driven building energy
forecasting model development, Energy and Buildings 183 (2019) 428–442.

[33] C. Deb, S. E. Lee, Determining key variables influencing energy consumption in office buildings through cluster analysis of pre-and post-retrofit building data, Energy and Buildings 159 (2018) 228–245.

[34] Z. Ma, P. Cooper, D. Daly, L. Ledo, Existing building retrofits: Methodology and state-of-the-art, Energy and buildings 55 (2012) 889–902.

[35] E. Kamel, A. M. Memari, Different methods in building envelope energy retrofit, in: 3rd Residential Building Design & Construction Conference, Penn State, University Park, 2016, pp. 2–3.

[36] D. Dagostino, P. Zangheri, L. Castellazzi, Towards nearly zero energy buildings in europe: A focus on retrofit in non-residential buildings, Energies 10 (1) (2017) 117.

[37] G. Liu, B. Liu, W. Wang, J. Zhang, R. A. Athalye, D. Moser, E. Crowe, N. Bengtson, M. Effinger, L. Webster, et al., Advanced energy retrofit guide office buildings, Tech. rep., Pacific Northwest National Lab.(PNNL), Richland, WA (United States) (2011).

[38] L. Jankovic, Lessons learnt from design, off-site construction and performance analysis of deep energy retrofit of residential buildings, Energy and Buildings 186 (2019) 319–338.

[39] 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, Applied Energy 205 (2017) 323–335.

[40] I. Ceballos-Fuentealba, E. Álvarez-Miranda, C. Torres-Fuchslocher, M. L. del Campo-Hitschfeld, J. Díaz-Guerrero, A simulation and optimisation methodology for choosing energy efficiency measures in non-residential buildings, Applied Energy 256 (2019) 113953.

[41] C. Delmastro, G. Mutani, S. P. Corgnati, A supporting method for selecting cost-optimal energy retrofit policies for residential buildings at the urban scale, Energy Policy 99 (2016) 42–56.

[42] J. S. Ramos, S. A. Domínguez, M. P. Moreno, M. G. Delgado, L. R. Rodríguez, J. A. T. Ríos, Design of the refurbishment of historic buildings with a cost-optimal methodology: A case study, Applied Sciences 9 (15) (2019) 3104.

[43] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: ACM sigmod record, Vol. 29, ACM, 2000, pp. 93–104.

[44] 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 (2018).
[55] U. Ali, M. H. Shamsi, C. Hoare, E. Mangina, J. O’Donnell, Application of intelligent algorithms for residential building energy performance rating prediction, in: 16th IBPSA International Conference Building Simulation, Italy, International Building Performance Simulation Association, 2019.

[56] T. A. Alamiedy, M. Anbar, A. K. Al-Ani, B. N. Al-Tanimi, N. Faleh, Review on feature selection algorithms for anomaly-based intrusion detection system, in: International Conference of Reliable Information and Communication Technology, Springer, 2018, pp. 605–619.

[57] F. Wang, Y. Yu, X. Wang, H. Ren, M. Shafie-Khah, J. Catalao, Residential electricity consumption level impact factor analysis based on wrapper feature selection and multinomial logistic regression, Energies 11 (5) (2018) 1180.

[58] Building Energy Rating Certificate Database by SEAI, [https://www.seai.ie](https://www.seai.ie) [Online; accessed 02-Mar-2020].

[59] Report on the Development of Cost Optimal Calculations and Gap Analysis for Buildings in Ireland under Directive 2010/31/EU on the Energy Performance of Buildings (RECAST), [https://www.housing.gov.ie/sites/default/files/publications/files/cost_optimal_residential_report_ireland_2018.pdf](https://www.housing.gov.ie/sites/default/files/publications/files/cost_optimal_residential_report_ireland_2018.pdf) [Online; accessed 2-Mar-2020] (2018).

[60] Nearly Zero Energy Building Standard by SEAI, [https://www.seai.ie/business-and-public-sector/standards/nearly-zero-energy-building-standard/](https://www.seai.ie/business-and-public-sector/standards/nearly-zero-energy-building-standard/) [Online; accessed 2-Mar-2020] (2019).

[61] J. Curtis, N. Devitt, A. Whelan, Using census and administrative records to identify the location and occupancy type of energy inefficient residential properties, Sustainable Cities and Society 18 (2015) 56–65.