Comparative analysis of prediction algorithms for building energy usage prediction at an urban scale

Usman Ali\textsuperscript{1}, Mohammad Haris Shamsi\textsuperscript{1}, Muhammad Nabeel\textsuperscript{2}, Cathal Hoare\textsuperscript{1}, Fawaz Alshehri\textsuperscript{1}, Eleni Mangina\textsuperscript{1}, James ODonnell\textsuperscript{1}

\textsuperscript{1}UCD Energy Institute, University College Dublin, Belfield, Dublin 4, Ireland
\textsuperscript{2}School of Systems and Technology, Department of Software Engineering, University of Management and Technology, Pakistan

E-mail: usman.ali@ucdconnect.ie

Abstract.
Strategic planning for efficient and sustainable urban environments necessitates identification of scalable energy saving opportunities for the buildings sector. A possible resolution is the analysis of building energy use data at urban scale, although the available data is often sparse, inconsistent, diverse and heterogeneous in nature. Over the past decades, predictive modeling using sparse data has aided with the forecasting of building energy use. However, most studies of energy use prediction focus on individual buildings. This paper proposes the integration of building archetypes simulation, parametric analysis, and machine learning techniques as a solution to accurately predict individual building energy use at an urban level. The aim of the research described in this paper is to achieve accurate prediction of building energy performance, which will allow stakeholders, such as energy policymakers and urban planners, to make informed decisions when planning retrofit measures at large scale. The methodology generates synthetic building data for training the predictive model and predicts building energy use at urban scale with limited resources. The experimentation focuses on Dublin city through the development of synthetic building dataset using parametric analysis on previously identified key variables of two distinct building archetypes. Having compared different prediction algorithms, we show that the Gradient Boosted Trees algorithm gives a better prediction when compared to other algorithms.

1. Introduction
The global final energy consumption contribution from buildings is more than 40%. At the same time, buildings are responsible for about 1/3 of greenhouse gas emissions \cite{2}. Urban planners and energy policymakers face significant challenges when identifying building energy performance opportunities in the context of strategic planning for efficient and sustainable urban environments. To cater for strategic sustainable planning, EU member states are establishing energy-efficient strategies to improve building energy performance. For instance, the Energy Performance of Buildings Directive (EPBD) introduced a new Directive (EU) 2018/844 that aimed to develop a sustainable and decarbonized energy system. The members of this directive must reduce their greenhouse gas emissions by 40% by 2030 compared to 1990 greenhouse gas emissions \cite{1}.

One possible solution to achieving higher sustainability and decarbonising the building stock is through retrofitting of existing buildings. To implement retrofit solutions on a large scale,
prior knowledge about the energy performance of existing buildings is often required. Urban planners and energy policymakers face a significant challenge as energy performance information is usually not available. At the same time, calculating building energy performance remains a challenging task due to numerous factors that affect the energy use such as building envelope, physical characteristics, installed heating-cooling equipment, climate conditions, and occupants' behavior.

Generally, there are two approaches - physical modeling and data-driven [3] - which have been applied for building energy usage calculation. Physical models (also known as engineering methods) are based on detailed building physics and are analyzed using simulation tools, for instance, EnergyPlus, ESP-r, and TRNSYS. These simulation tools require detailed input with geometric and non-geometric information of any building and failure to provide accurate inputs can produce incorrect results. Similarly, to simulate an entire district, a massive amount of data is required for each building. On the other hand, a data-driven approach predicts energy usage based on historical data and use of machine learning algorithms. This approach does not require detailed knowledge about the building when compared to the physical modeling approach. As data-driven approaches can predict the energy consumption with limited available information, these approaches have gained a lot of attention in the energy sector during the past few years [3]. Furthermore, these approaches often tend to obtain the highest levels of accuracy as the models are individually trained with the building energy usage data.

Generally, machine learning algorithms required for the data-driven approach are further divided into two main categories such as classification algorithms and regression [3]. A classification algorithm is used when the output variable is denoted as a designated label, such as energy rating or building type. Some common classification algorithms include nearest neighbor, naive Bayes, rule induction, deep learning, Support Vector Machines (SVM) and neural networks, etc [8]. Regression algorithms are used when the output variable is a real value such as energy consumption. The most common regression algorithms include generalized linear models, deep learning, decision tree, random forest, gradient boosted trees and support vector machine [8].

Mostly, studies using the data-driven approach focus on a single building energy use prediction [7]. One of the main reasons is the lack of high-quality and useful data at a large scale. Few studies focus on large scale energy prediction; for instance, Tardioli et al. proposed a methodology to generate a buildings database of 4096 buildings using parametric simulation and then applied the prediction algorithms to generate building load profiles [7]. However, the study used only a few parameters for data generation and the database lacked crucial parameters such as roof U-values, type of HVAC systems, presence of renewable energy systems etc.

This paper introduces a methodology to facilitate the prediction of building energy use at an urban scale using synthetic data. The main objective of this paper is to generate synthetic data using parametric simulation on previously identified key variables and compare prediction algorithms that can be used to predict building energy use at an urban scale.

The paper is organized as follows: Section 2 provides a detailed discussion of the devised methodology for residential building energy use prediction. Section 3 discusses the case study of the Irish building stock and compares different machine learning algorithms in term of prediction performance. Finally, conclusions are discussed in Section 4.

2. Methodology
The devised methodology in this research uses machine leaning algorithms to predict the building energy consumption (Figure 1). The methodology starts with archetypes development followed by parameter selection, parametric simulation, implementation of prediction algorithms and ends with model performance analysis.
In a large building stock, several buildings often possess similar characteristics and hence, can be represented by building archetypes. In the parametric simulation framework, each building archetype is used as a base model. The building geometric and non-geometric data is required for simulation of any building archetype. The building geometric data includes shapes and proportions of the building. The non-geometric data includes commonly used physical properties, heating/cooling system etc. These data can be extracted from existing available building stock databases, for instance, TABULA [5]. Finally, weather data sets are also required for building thermal simulations.

2.2. Parameter Selection
The selection of parameters needed to perform parametric simulation are a part of the synthetic data set and is a crucial step because it will affect the overall accuracy of the prediction model. The parameter values include all the variations required for synthetic data generation. The selection of key parameters and their variations can be easily found in existing literature or surveys of specific climate environment or area [4, 6]. Common construction parameters include properties of roof, wall, floor, ceiling and window. Internal gains, occupancy density, heating and cooling systems are also key parameters that are used in the parametric simulation.

2.3. Parametric Simulation
In this paper, JEPlus is used as a parametric tool for energy simulations. JEplus uses EnergyPlus for simulation and design builder templates for integration of different parameter values. Due to the complexity introduced by a large number of parameters, it is almost impossible to generate the simulated data for all parameters. Therefore, synthetic data is extracted by sampling methods such as Simple Random Sampling (SRS) and Latin Hypercube Sampling (LHS). These methods help in generating desired sample data that contain variation of all parameters.

2.4. Predictive Model Adaptation
In the predictive model adaptation process, the first step involves pre-processing of the synthetic data to prepare the input of the prediction algorithms. The data is split into two subsets; a training set (a subset to train a model) and a test set (a subset to test the trained model). Generally, data splitting applies one of two techniques, random data splitting and cross-validation. In random data splitting, the random data is split into training and test sets, according to a 80/20 split ratio respectively. Cross-validation is the most common technique to achieve a balance between minimal bias and variance of the training model. In cross-validation, the data is first divided into k subsets and then the data splitting process is applied to each
sub-set. In each $k^{th}$ iteration, a different subset is used for testing while the other (k-1) sets are used for training. In this paper, the cross-validation algorithm is used for data splitting.

Regression algorithms are used to predict the value of given set of data points. In this paper, six different learning algorithms are used for rating such as Generalized Linear Model (GLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT) and Support Vector Machine (SVM). GLM is a flexible generalization of conventional linear regression models for a continuous response variable given continuous predictors. DL also known as deep structured learning, uses neural network architectures. DT builds regression models as tree-like structures in which each node represents a splitting rule for one specific attribute. RF is an ensemble of decision trees at training time and gives output mean prediction (regression) of the individual trees. GBT is also an ensemble of regression based tree models, that uses a flexible nonlinear regression procedure using forward-learning ensemble methods to improve the accuracy of trees. SVM is a discriminative algorithm. In other words, given a set of real value training data, the algorithm outputs a model that assigns existing categories to the new examples, thus, making a non-probabilistic binary linear classifier. These six algorithms offer excellent performance when used for energy forecasting as evident from previous studies [8, 3].

2.5. Model Performance Analysis

To examine the effectiveness of the learning prediction models, adopted performance indices such as Root Mean Squared Error (RMSE) and Absolute Error (AE) are used [8]. An algorithm exhibiting the lowest RMSE and AE values is considered to be the best amongst all algorithms.

3. Case Study

The main objective of this paper is to develop an energy use prediction system for energy policymakers or urban planners. The proposed methodology is applied to the Irish residential building stock. The experiment focuses on Dublin city through the development of synthetic building dataset using parametric simulation on previously identified key variables of two distinct building archetypes. In 2016, there were more than 1,983,715 residential buildings in Ireland.

In this paper, we performed the analysis on semi-detached and apartment building archetypes. In 2016, 28% of the buildings in Ireland were semi-detached while 12% were apartments.

To perform a parametric simulation with the archetypes, we first identified the non-geometric and geometric parameters associated with the existing building stock of Dublin city. The commonly used non-geometric parameters are identified based on existing literature. For instance, studies by Egan et al. have identified non-geometric parameters that are relevant and would influence the energy performance of the Irish building stock [4]. We extended the list of parameters identified by these studies to perform a detailed HVAC system analysis. Some of the parameters added were HVAC systems, fuel type and renewable parameters. Parameters needed for parametric simulation of archetypes are listed in Table 1. Similarly, geometric information of archetypes is also collected from existing literature by Egan et al [4].

We implemented the LHS method to generate the sample data of 20,000 buildings for predictive algorithms. The data is split into two parts to create training and testing data using a cross-validation algorithm.

The synthetic data is trained and compared using six different algorithms and a robust model input is chosen to evaluate the sensitivity of the learning algorithms. In GLM, the algorithm run on the operator starts a one-node local H2O cluster. For the DL algorithm, we used two hidden layers of size 50 each. The dataset is iterated ten times to achieve the best results. In the DT algorithm, the least square criteria is used for splitting the branch. In RF, we used an ensemble of 60 random trees with the least square criteria. Similarly, for the GBT algorithm, 140 trees are used to form an ensemble of regression-based tree models. Finally, for the SVM algorithm,
Table 1: Parameters needed for parametric simulation of archetypes

| Number | Parameters                          | Unit                        |
|--------|------------------------------------|-----------------------------|
| P1     | Building archetypes                |                             |
| P2     | Wall U-value                       | W/m²K                       |
| P3     | Window U-value                     | W/m²K                       |
| P4     | Floor U-value                      | W/m²K                       |
| P5     | Roof U-value                       | W/m²K                       |
| P6     | Door U-value                       | W/m²K                       |
| P7     | Orientation                        | ◦                           |
| P8     | Lighting power density             | W/m²-100lux                 |
| P9     | Occupancy density                  | Person(s)                   |
| P10    | Equipment density                  | W/m²                        |
| P12    | Heating setpoint temperature       | °C                          |
| P13    | HVAC system                        |                             |
| P14    | HVAC efficiency                    | %                           |
| P15    | HVAC fuel type                     |                             |
| P16    | Renewables                         | boolean                     |

The mySVM java implementation is used with a radial kernel type. The best algorithm with minimum RMSE and AE values for each archetype is shown in Figure 2.

![Figure 2: Comparison of prediction algorithms performance for semi-detached and apartment archetypes of Dublin city.](image)

The results show that Gradient Boosted Trees algorithm performed best for both archetype buildings because it handles missing data. The quality of EPC data used for training was not good and boosting proved effective when dealing with unbalanced data. The lowest RMSE achieved for the semi-detached archetype was 11.50 while that for apartments was 5.79. Similarly, the lowest AE achieved for the semi-detached archetype was 7.84 while that for apartments was 3.97. The results indicate that the same algorithm gives better results for both archetypes even though a significant differences exists between the two archetypes in terms of their physical structure and parameter values. The gradient boosted trees algorithm is able to capture the complex relationship between energy consumption values and building site specifications in the most efficient manner.
4. Conclusions and Future Work
Building energy use analysis at a large scale helps the stakeholders to formulate policy measures for reducing energy consumption and CO\textsubscript{2} emissions. The collection and identification of building energy performance data is a complex and time-consuming task, where numerous resources are required for such data collection at an urban scale. The devised methodology in this work implements machine learning algorithms to identify building energy use for an entire urban building stock. The proposed methodology is able to predict a buildings’ current energy use even with limited knowledge of the building dynamics.

The results presented focus on Dublin city through the development of synthetic building data set using parametric analysis on previously identified key variables of two distinct building archetypes. The comparison between different machine learning algorithms shows that the Gradient Boosted Trees algorithm gives a better prediction when compared to other algorithms. The accurate prediction of building energy performance will allow stakeholders such as energy policymakers and urban planners to make informed decisions when planning retrofit measures at large scale.

As the proposed methodology is only tested with EnergyPlus software, variations might occur with other software. The performance of the proposed best algorithm might turn out to be different for other areas or countries because of different building structures and climate environments. However, the methodology to predict energy use with synthetic data or parameters is consistent for an urban scale.

Further test cases could involve the validation of the methodology for areas with different building structures and climate types and at different scales. Future work will also consider cloud computing parametric simulation and application of hybrid approaches for different prediction algorithms.

Acknowledgments
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. The opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Science Foundation Ireland.

References

[1] European Union Directive (EU) 2018/844 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.

[2] EU-Energy 2018 Energy for Europe by European Commission https://ec.europa.eu/energy/en [Online; accessed 15-May-2019].

[3] Amasyali K and El-Gohary NM 2018 A review of data-driven building energy consumption prediction studies Renewable and Sustainable Energy Reviews 81:1192-205.

[4] Egan J, Finn D, Soares PH, Baumann VA, Aghamolaei R, Beagon P, Neu O, Pallonetto F and O’Donnell J 2018 Definition of a useful minimal-set of accurately-specified input data for building energy performance simulation Energy and Buildings 165:172-83.

[5] Loga T, Stein B, Diefenbach N.2016 TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable Energy and Buildings 132:4-12.

[6] Kapetanakis DS, Mangina E, Ridoluce EH, Kouromas K and Finn D. 2015 Selection of Input Variables for a Thermal Load Prediction Model Energy Procedia 78:3001-6.

[7] Tardioli G,Kerrigan R, Oates M, O’Donnell J, and Finn D. 2017 A data-driven modelling approach for large scale demand profiling of residential buildings IBPSA Building Simulation Conference.

[8] Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, Han M and Zhao X 2018 A review of data-driven approaches for prediction and classification of building energy consumption Renewable and Sustainable Energy Reviews 82:1027-47.