Tuning Machine Learning Models for Prediction of Building Energy Loads

Saleh Seyedzadeh[a,c,*], Farzad Pour Rahimian[b], Parag Rastogi[c], Ivan Glesk[a]

[a] *Faculty of Engineering, University of Strathclyde, Glasgow, UK.*
[b] *Faculty of Engineering & Environment, Northumbria University, Newcastle, UK.*
[c] *arbnco, Glasgow, UK.*

Abstract

There have been numerous simulation tools utilised for calculating building energy loads for efficient design and retrofitting. However, these tools entail a great deal of computational cost and prior knowledge to work with. Machine Learning (ML) techniques can contribute to bridging this gap by taking advantage of existing historical data for forecasting new samples and lead to informed decisions. This study investigated the accuracy of most popular ML models in the prediction of buildings heating and cooling loads carrying out specific tuning for each ML model and using two simulated building energy data generated in EnergyPlus and Ecotect and compared the results. The study used a grid-search coupled with cross-validation method to examine the combinations of model parameters. Furthermore, sensitivity analysis techniques were used to evaluate the importance of input variables on the performance of ML models. The accuracy and time complexity of models in predicting heating and cooling loads are demonstrated. Comparing the accuracy of the tuned models with the original research works reveals the significant role of model optimisation. The outcomes of the sensitivity analysis are demonstrated as relative importance which resulted in the identification of unimportant variables and faster model fitting.

Keywords: Building energy loads, Energy prediction, Machine learning,

*Email: s.seyyedzadeh@gmail.com,
Address: 605, 75 Montrose St, Glasgow G1 1XJ

Energy modelling, Energy simulation, Building design

1. Introduction

Buildings must be designed to maximise the health and well-being of their occupants while consuming the least energy and materials possible. Improving the building stock to achieve this requires improvements to existing buildings in addition to the construction of new high-performance buildings. One approach to the design of high-performance buildings is performance-driven design, in which the energy used by a building to keep its occupants comfortable is approximated using a physics-based simulation program. This method is called Building performance Simulation (BPS), and it allows a designer/engineer to examine the influence of form, materials, and systems before construction on the expected thermal performance of a building. Conventionally, the search for an optimal design with simulation has been through a manual iterative process - design, analyse, change. This cycle is a labour-intensive task, so the search space, i.e., the scope of possible options, is necessarily limited. The use of performance-driven design can, thus, be augmented with optimisation, since optimisation offers a way to significantly expand the search space during the design process.

The benefits of optimisation over manual search are realised when the optimising routine is able to evaluate thousands of potential options (Si, 2017). However, large runs of performance simulations of realistic building models require significant time and computational resources. Optimisation reduces the specialist labour required to search very large spaces of options, but the resulting computational load can overwhelm the design process. The use of surrogate models has been proposed to overcome this problem (Zhao & Magoulès, 2012a). Surrogate or data-driven models are mathematical relationships between inputs and outputs of interest from the system being studied, learnt from measured or simulated data that represents the physical problem. For example, the thermophysical properties of building materials and weather parameters can be used

to predict indoor environmental conditions, as we do in this paper. Sufficiently precise surrogate models, thus, provide fast and accurate alternatives to building performance simulators during a computationally-intensive design process (Rastogi et al., 2017).

The use of surrogate models requires careful consideration of the accuracy and appropriateness of the data and relationships inferred from the data. In this paper, we examine a practical aspect of this approach: selecting and tuning regression models for a given dataset. By this, we mean selecting the model types, structures, and parameters most appropriate to the problem at hand. As we have described in the next section, most previous work in using surrogate models in building simulation either compares linear models with nonlinear models or different types of nonlinear models (Seyedzadeh et al., 2018). In addition, previous work has usually only optimised a limited number of model parameters. The selection of model parameters, however, determines the performance of a model on a given dataset, and this performance varies from one dataset to another. Thus, the previous work does not provide a complete evaluation of different nonlinear models and does not provide sufficient guidance about model selection. We show that the process of selecting a model must account not just for predictive accuracy but also model complexity, ease of use, and consistency of predictions. We use the datasets described by Rastogi et al. (2017) and Tsanas & Xifara (2012) to demonstrate the performance of different candidate models.

The paper is organised as follows. The next section presents a review of previous studies and issues with using ML models in predicting building energy consumption. That is followed by a description of the nonlinear methods evaluated, the case studies, and results from our tuning proposals. The final section contains recommendations on model selection and discusses future work.

2. Background and Motivation

This paper addresses the improvement of regression algorithms that relate building characteristics with performance indicators of interest, e.g, insulation level of walls with energy used for space heating. Broadly, regression approaches are divided into two categories: supervised learning, in which the target is known, and unsupervised learning, where there is no "output" to learn and predict. A supervised learning problem is either one of regression or classification. In both cases, input features (X) are mapped to one or more output variables (Y), such that changes in inputs cause the changes in output expected from the real system being modelled. Unsupervised learning includes techniques such as clustering, which organises data into groups based on similarities among the samples in a dataset. Unsupervised learning is applied to an unlabelled dataset, i.e., where the there are no labels or target values for the model to learn from, while a supervised learning algorithm can be tested against some 'known' labels or values. In the use case demonstrated here, we have a database of inputs and outputs obtained from a physics-based simulator, so this is a supervised learning exercise. This simulator is the 'ground truth', and the models are judged solely on their ability to represent the simulator. We show how a model may be improved to better represent the behaviour of a simulator while providing estimates of outputs in a fraction of the time it would take the simulator to run a simulation.

We will focus on nonlinear regression models, i.e., where the inputs cannot be combined linearly, with or without any transformations. For example, a polynomial model can be reformulated as a linear model of transformed (squared, cubed, etc.) inputs but no such transformations can be applied to the inputs of a Random Forest (RF) regression model. The use of machine learning (ML) models in the analysis of buildings was first demonstrated by Kalogirou et al. (1997) to estimate building heating loads considering envelope characteristic and the desired (setpoint) temperature. Since then, the studies described below have demonstrated that nonlinear models predict both simulated and metered

data better than linear models. Several studies have also presented comparisons between different nonlinear models. We now present an overview of the literature, organised by the type of model(s) used in each study.

The pioneering work of Kalogirou et al. (1997) was completed in 2000 by using Artificial Neural Networks (ANN) to predict the hourly energy demand of holiday dwellings. Kalogirou et al. (2001) also used ANN to estimate the daily heat loads of model house buildings with different combinations of the wall and roof types (e.g., single vs cavity walls and roofs with different insulation) using typical meteorological data for Cyprus. In that study, TRNSYS was used to estimate energy use and the estimates were validated by comparing one building with actual measurements. Similarly, Yokoyama et al. (2009) used a global optimisation method coupled with the ANN to predict cooling load demand. The authors probed two parameters of the network, namely the number of hidden layers and the number of neurons in each layer.

Paudel et al. (2014) incorporated occupancy profiles as well as features representing the operational heating power level and climate variables to model the heating energy consumption. To increase the accuracy of the prediction model, time-dependent attributes of operational heating power level were also included. Deb et al. (2016) used five days' data as inputs to forecast the daily cooling demand of three institutional buildings in Singapore. Mena et al. (2014) applied ANN for short-term forecasting of hourly electricity consumption using time series of solar production over two years in Spain. The resulting model highlighted the significant impact of outdoor temperature and solar radiation on electricity usage. Platon et al. (2015) used Principal Component Analysis (PCA) to explore the selection of inputs for an ANN model in the prediction of hourly electricity consumption of an institutional building. Li et al. (2015) also used PCA to reduce the number of input variables, and introduced a new optimisation algorithm to improve the performance of an ANN model for short-term forecasting of electricity demand. Neto & Fiorelli (2008) applied ANN to predict the daily metered energy usage of a commercial building and demonstrated that supervised learning could produce more accurate predictions than EnergyPlus

software. Improper assessment of lighting and occupancy was recognised as the primary source of model uncertainty. Dombayci (2010) used degree-hours, i.e., the difference between the external temperature and some nominal balance point temperature integrated over time, to estimate the hourly energy demand and used that as the output to be predicted. Due to its use of only degree-hours the proposed approach was accurate only for the simplest residential buildings.

Yalcintas (2006) applied ANN to approximate the energy performance of sixty educational buildings in Hawaii. The data was collected from energy assessments reports and contained information about buildings and their air conditioning systems. Wong et al. (2010) estimated the dynamic energy and daylighting performance of a commercial building. The building daily energy usage was calculated using EnergyPlus and coupled with an algorithm to approximate the interior reflections. Hong et al. (2014b) found that statistical analysis was less accurate than ANNs in evaluating the energy performance of schools in the UK. Khayatian et al. (2016) applied ANN to predict energy performance certificates of domestic buildings in Italy. Ascione et al. (2017) investigated the association of energy usage and occupant thermal comfort in the prediction of energy performance. The energy consumption was calculated using EnergyPlus, and the ANN model trained on these was used as the calculation engine to optimise building design and retrofit planning.

The use of Support Vector Machines (SVM) for forecasting building energy consumption was introduced by Dong et al. (2005). The model was trained with temperature, humidity, and solar radiation as the inputs. Li et al. (2009) and Hou & Lian (2009) used SVM to forecast hourly cooling leads of an office building considering the same parameters suggested by Dong et al. (2005). Massana et al. (2015) also investigated the short-term prediction of electricity consumption using SVM. Xuemei et al. (2009) improved the performance of SVM used for predicting cooling loads. Li et al. (2010) applied SVM for yearly estimation of electricity consumption in domestic buildings. The model considered building envelope parameters as well as the annual electricity consumption normalised by unit area. Zhao & Magoulès (2010) applied a parallel imple-

mentation of SVM to calculate energy usage for optimising the design of office buildings. The work was improved by reducing the input variable space by applying gradient-guided feature selection and the correlation coefficients method. Jain et al. (2014) investigated the impact of different time intervals and building spaces in data collection on energy demand forecasting using SVM. Finally, Chen & Tan (2017) used an SVM model coupled with multi-resolution wavelet decomposition for estimating energy consumption of various building type.

Since early 2000, Gaussian Process (GP) regression has been used by researchers in different applications, especially where there are uncertainties in input parameters (Grosicki et al., 2005; Bukkapatnam & Cheng, 2010). Heo et al. (2012) and Heo & Zavala (2012) used GP modelling to estimate uncertainty levels in calculating building energy saving after retrofitting. Zhang et al. (2013) also applied GP to predict the post-retrofit energy demand of an office building. Burkhart et al. (2014) incorporated GP with a Monte Carlo expectation maximisation algorithm to train the model under data uncertainty, to optimise the performance of the HVAC system of an office building. They found that the models can be trained even with limited data or sparse measurements when precise and extensive sensor data is not available. Rastogi et al. (2017) compared the accuracy of GP and linear regression in emulating of a building performance simulation and show that the accuracy of GP is four times better than linear regression.

Though ensemble ML models such as Random Forest (RF) and Gradient Boosted Regression Trees (GBRT) have been around for decades, their use in estimating building energy consumption is very new. Tsanas & Xifara (2012) applied RF to estimate energy consumption using building characteristics and showed that it performed better than the iteratively re-weighted least squares method. Papadopoulos et al. (2017) compared tree-based models in building energy performance estimation using the data provided by Tsanas & Xifara (2012). Recently, Wang et al. (2018) used RF for short-term prediction of energy usage in an office building considering compound variables of envelope, climate, and time.

Yalcintas & Ozturk (2007) and Yalcintas (2006) compared the accuracy of Artificial Neural Networks (ANN) with that of Multiple Linear Regression (MLR) to estimate the energy performance of commercial buildings, while Hong et al. (2014a) compared ANN with physics-based models for the same use case. Platon et al. (2015) compared ANN with case-based reasoning (CBR) for prediction of hourly electricity consumption of an institutional building. Li et al. (2009) applied SVM to forecast hourly cooling loads of an office building and provided a comparison with ANN, indicating that the ANN had more potential. However, Edwards et al. (2012) also evaluated the accuracy of SVM and ANN in forecasting hourly energy consumption of residential buildings and found ANN to be the least accurate model. Zhang et al. (2015) compared change-point models with Gaussian-based and one layer ANN models to predict the energy consumed by an office building to supply hot water and concluded that ANN models can be inefficient when enough data is not available. Tsanas & Xifara (2012) investigated the accuracy of RF and iteratively re-weighted least squares regression model. Wang et al. (2018) indicated the superiority of RF over SVM and regression trees in predicting hourly building energy demands.

2.1. Gaps

As the survey of literature in this section shows, there is a wealth of examples of nonlinear regression models (machine learning) applied to problems related to predicting energy and mass flows in buildings. Each study demonstrates the use of one model type/architecture or comparison between different model types. However, there is a lack of guidance on how to optimise or 'tune' models to fit the problem at hand for the best predictive accuracy and consistency. This paper, therefore, lays out a widely-applicable approach to tuning nonlinear models to building energy data. Though the examples shown here are from simulators, the method is also applicable to measured data.

3. Machine Learning Models

ML models operate as a black box, so further information about the building is not required. The general scheme of supervised learning for modelling building energy is illustrated in Figure 1. As seen, the first step is to select a set of features for representing the building energy system. Although data-driven methods build models with fewer variables than engineering techniques, it is crucial to generate a logical input set for the ML model. These features are not necessarily raw building characteristics or weather data; instead, they could be complex variables calculated from basic ones, e.g. wall to floor ratio and mean daily global radiation (Zhao & Magoulès, 2012b).

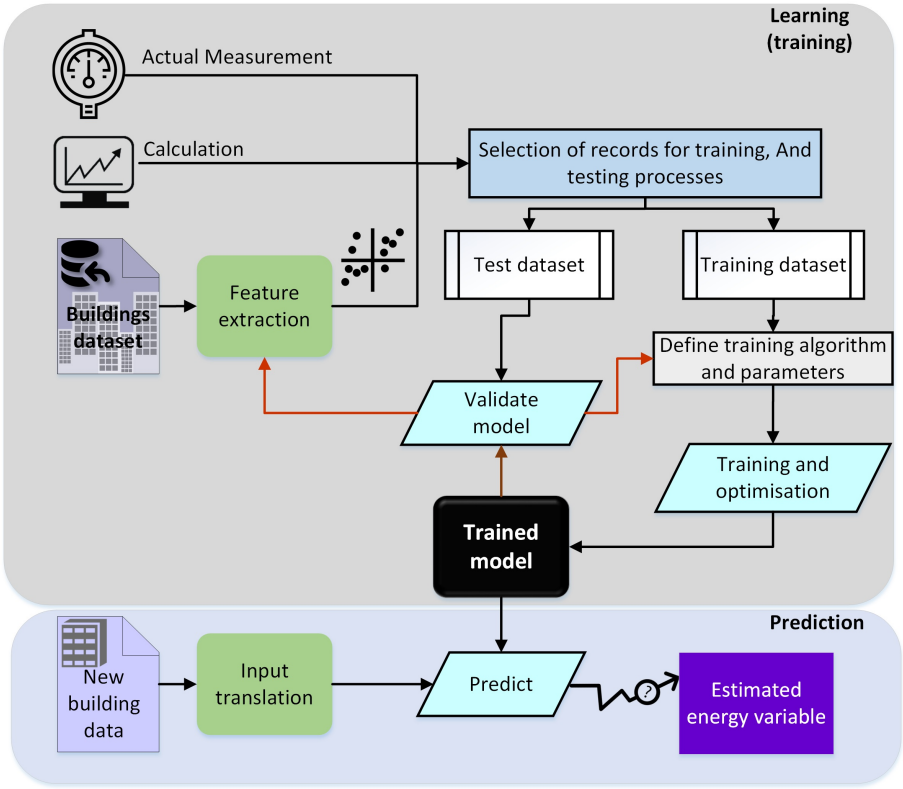


Figure 1: General schematic diagram of supervised learning.

The next key stage in utilising MLs is optimisation of model itself. This

procedure which is called tuning plays an important role in the performance of an ML model especially when it is a complex one. Choosing inappropriate hyper-parameters will result in poor accuracy which may falsely be translated as the model failure. Although selecting the right input variables is essential for training a successful machine, the full advantage cannot be taken of ML without tuning the model for that specific training data. Each ML has different hyper-parameters which govern the learning process. A key point in tuning an ML model parameters is the generalisation. That is to say, how well the learning model applies to specific examples not seen by the model when it was training. Hence, in the procedure of model optimisation, there should be a proper mechanism such as cross-validation to avoid overfitting (i.e. modelling the training data too well).

Five ML techniques including ANN, SVM, GP, RF and GBRT are employed to emulate two BPS tools namely EnergyPlus and Ecotect. A standard method that we used to select optimal hyper-parameters is a grid-search combined with k-fold cross-validation. In this procedure, the data is divided into k exclusive subsets, and each combination of model parameters and architecture is fitted to each distinct group of $k-1$ subsets and tested on the remaining subset. This process provides a distribution of errors for a given model choice on different parts of the dataset, i.e., an estimate of the general applicability of the model to represent the variation in the dataset. Furthermore, different normalisations such as standard, min-max and robust are applied to data before training procedure. Robust scaler eliminates the median and normalises data according to the inter-quartile range.

Basics of each model and the parameters going under optimisation are explained as followings.

3.1. Artificial Neural Network

Neural networks have been broadly utilised for building energy estimation and known as the major ML techniques in this area. They have been successfully used for modelling non-linear problems and complex systems. By applying

different techniques, ANNs have the capability to be immune to the fault and noise (Tso & Yau, 2007) while learning key patterns of building systems.

The main idea of ANN is obtained from the neurobiological field. Several kinds of ANN have been proposed for different applications including, Feed Forward Network (FFN), Radial Basis Function Network (RBFN) and recurrent networks (RNN). Each ANN consists of multi-layers (minimum two layers) of neurons and activation functions that form the connections between neurons. Some frequently used functions are linear, sigmoid ad hard limit functions (Park & Lek, 2016).

In FFN which was the first NN model as well as the simplest one, there are no cycles from input to output neurons and the pieces of information moves in one direction in the network. Figure 2 illustrates the general structure of FFN with input, output and one hidden layer.

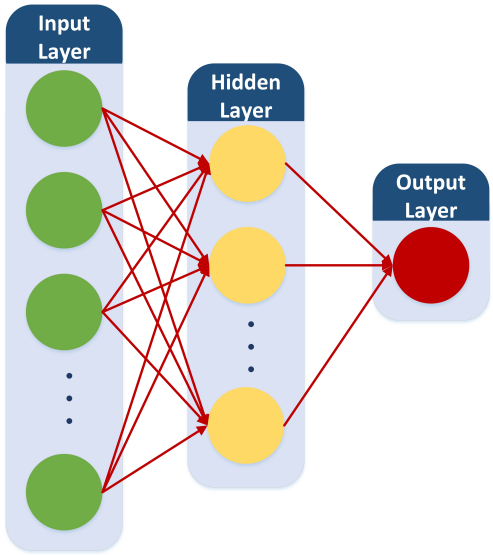


Figure 2: Conceptual structure of feed forward neural network with three layers.

RNN uses its internal memory to learn from preceding experiences by allowing loops from output to input nodes. RNN is proposed in various architectures including fully connected, recursive, long short-term memory, etc. This type

of neural network has usually been employed to solve very deep learning tasks such as multivariate time-series prognostication; often more than 1000 layers are needed (Ghiassi et al., 2005).

In RBFM, a radial basic function is used as activation function providing a linear combination of inputs and neuron parameters as output. This type of network is very effective for time series estimation (Harpham & Dawson, 2006; Leung et al., 2001; Park et al., 1998).

Due to the nature of the datasets, a multilayer perception FFN is utilised in this work. The ANN hyper-parameters which go under optimisation are:

- **Optimiser:** the function that updates the weights and bias;
- **Activation:** a non-linear transformation function which is applied over the input, and then the output is fed to the subsequent layer neurons as input. An ANN without activation function will act as a linear regressor and may fail to model complex systems;
- **Initialisation:** the initial values of weights before the optimiser is applied for training;
- **Epoch:** the number of forward and backward passes for all samples of data;
- **Batch size:** specifies the number of samples that are propagated through the ANN training (i.e. the number of samples in one epoch);
- **Dropout rate:** dropout is a regularisation method for preventing ANN from overfitting and creating more generalised model by randomly rejecting some neurons during training. Droput rate determines the percentage of randomly input exclusion at each layer;
- **Size:** number of neurons in each layer and number of layers.

3.2. Support Vector Machine

SVMs are highly robust models for solving non-linear problems and used in research and industry for regression and classification purposes. As SVMs can

be trained with few numbers of data samples, they could be right solutions for modelling study cases with no recorded historical data. Furthermore, SVMs are based on the Structural Risk Minimisation (SRM) principle that seeks to minimise an upper bound of generalisation error consisting of the sum of training error and a confidence level. SVMs with kernel function acts as a two-layer ANN, but the number of hyper-parameters is fewer than that. Another advantage of SVM over other ML models is uniqueness and globally optimality of the generated solution, as it does not require non-linear optimisation with the risk of sucking in a local minimum limit. One main drawback of SVM is the computation time, which has the order almost equal to the cube of problem samples.

Suppose every input parameter comprises a vector X_i (i denotes the ith input component sample), and a corresponding output vector Y_i that can be building heating loads, rating or energy consumption. SVM relates inputs to output parameters using the following equation:

$$Y = W \cdot \phi(X) + b \tag{1}$$

where $\phi(X)$ function non-linearly maps X to a higher dimensional feature space. The bias, b, is dependent of selected kernel function (e.g. b can be equal to zero for Gaussian RBF). W is the weight vector and approximated by empirical risk function as:

$$Minimise : \frac{1}{2}\|W\|^2 + C\frac{1}{1}\sum_{i=1}^{N} L_\varepsilon(Y_i, f(X_i)) \tag{2}$$

L_ε is ε-intensity loss function and defined as

$$L_\varepsilon(Y_i, f(X_i)) = \begin{cases} |f(x) - Y_i| - \varepsilon, & |f(x) - Y_i| \geq \varepsilon \\ 0, & otherwise \end{cases} \tag{3}$$

Here ε denotes the domain of ε-insensitivity and N is the number of training samples. The loss becomes zero when the predicted value drops within the band area and gets the difference value between the predicted and radius ε of the

domain, in case the expected point falls out of that region. The regularised constant C presents the error penalty, which is defined by the user.

SVM rejects the training samples with errors less than the predetermined ε. By acquisition slack variables ξ and ξ_i^* for calculation of the distance from the band are, equation (3) can be expressed as:

$$\underset{\xi,\xi_i^*,W,b}{Minmise} : \frac{1}{2}\|W\|^2 + C\frac{1}{N}\sum_{i=1}^{N}\xi + \xi_i^* \tag{4}$$

subject to

$$\begin{cases} Y_i - W \cdot \phi(x_i) - b \leq \varepsilon + \xi \\ W \cdot \phi(x_i) + b - Y_i \leq \varepsilon + \xi_i^* \\ \xi \geq 0, \quad \xi_i^* \geq 0 \end{cases} \tag{5}$$

The SVM problem using a kernel function of $K(X_i, X_j)$ (α_i, α_i^* as Lagrange multipliers) can be simplified as:

$$\begin{aligned} \underset{\{\alpha_i\},\{\alpha_i^*\}}{Maximise} : -\varepsilon\sum_{i=1}^{N}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} Y_i(\alpha_i^* - \alpha_i) - \\ \frac{1}{2} sum_{i=1}^{N} \sum_{j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(X_i, X_j) \end{aligned} \tag{6}$$

subject to

$$\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0, \qquad 0 \leq \alpha_i, \alpha_i^* \geq C \tag{7}$$

As mentioned before the number of parameters in SVM with a Gaussian RBF kernel is few as two which are C and Gamma.

3.3. Gaussian Process

The main drawback of GP modelling is expensive computational cost, especially with the increase of training samples. This is due to the fact that GP

constructs a model by determining the structure of a covariance matrix composed of $N \times N$ input variable where the matrix inversion required in predictions has a complexity of $O(N^3)$

Given a set of n independent input vector X_j $(j = 1, \cdots, n)$, the corresponding observations of y_i $(i = 1, \cdots, n)$ are correlated using covariance function K with normal distribution equal to (Li et al., 2014):

$$P(y; m; k) = \frac{1}{(2\pi)^{n/2}|K(X,X)|^{1/2}} \times exp\left(-\frac{1}{2}(y-m)^T K(X,X)^{-1}(y-m)\right) \tag{8}$$

The covariance or kernel function can be derived as

$$K = \begin{vmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{vmatrix} \tag{9}$$

A white noise, σ, is presumed in order to consider the uncertainty. It is assumed that the samples are corrupted (lets suppose as new inputs as x^*) by this noise. In this case covariance of y is expressed as

$$cov(y) = K(X,X) + \sigma^2 \tag{10}$$

Then y^* can be estimated as below.

$$y^* = \sum_{i=1}^{n} \alpha_i k(x_i, x^*) \tag{11}$$

$$\alpha_i = \left(K(X,X) + \sigma^2 I\right)^{-1} y_i \tag{12}$$

For GP model three parameters are tuned: kernel, alpha (α) which is the value added to the diagonal of the kernel matrix (equation 11) and the number of restarts of the optimiser for discovering the parameters maximising the log-

marginal probability. Two combinations of white noise with RBF and Matern covariance functions are used for GP model kernel. Matern kernel is denied as:

$$K(X, X') = \frac{261 - v}{\Gamma(v)} \left(\frac{\sqrt{2v} \mid x - x' \mid}{I} \right)^{v} K_v \left(\frac{\sqrt{2v} \mid x - x' \mid}{I} \right) \tag{13}$$

Here, Γ is the Gamma function and K_v is the modified Bessel function the second-order v (Owen et al., 1965).

3.4. Random Forest

Random forest is a collection (ensemble) of randomised decision trees (DTs) (Tin Kam Ho, 1995). DT is a non-parametric ML that establishes a model in the form of a tree structure. DT repeatedly divides the given records into smaller and smaller subsets until only one record remains in the subset. The inner and final sets are known as nodes and leaf nodes. As the precision of DT is substantially subject to the distribution of records on in the learning dataset, it is considered as an unstable method (i.e. tiny alteration in the observations will change the entire structure). To overcome this issue a set of DTs and uses the average predicted values of all independent trees as the final target. In general, RF applies bagging to combine separate models but with sore of similar information and generate a linear combination from many independent trees.

RF requires few number of hyper-parameters to be set. The main parameter is the number of independent trees in the forest. There is a trade-off between the accuracy of model and training/prediction computational cost. Thereby, this parameter should be tuned to choose the optimal value. Other parameters include the number of features to consider when seeking for the best split, whether bootstrap samples are used when creating trees and minimum number of data sample to split a node and required in each node.

3.5. Gradient Boosted Regression Trees

Like RF, GBRT is an ensemble of other prediction models such as DTs. The principal difference between GBRT and RF is that the latter one is based on fully developed DTs with low bias and high variance, while the former employs weak

learners (small trees) having high bias and low variance (Breiman, 2017). In GBRT, trees are not independent of each other; instead, each branch is created based on former simple models through a weighting procedure. This approach is known as boosting algorithm. At each inner node (i.e. the split point) given dataset is divided into two samples. Let's assume a GBRT with three nodes trees; then there will be one split point in which the best segmentation of the data is decided, and the divergence of the obtained values (from the individual averages) are calculated. By fitting on these residuals, the subsequent DT will seek for another division of data to reduce the error variance.

Most important parameters for optimising GBRT comprise learning rate (also known as shrinkage) which is a weighting procedure to prevent overfitting by controlling the contribution of each tree, number of trees, maximum depth of tree and the number of features for searching best division, and the minimum number of data sample to split a node and required in each node. Moreover, the sub-sample parameter defines the fraction of observation to be selected for each tree.

Rather than conventional GBRT model the recently improved version known as eXtreme Gradient Boosting (XGBoost) algorithm (Chen & Guestrin, 2016) is also evaluated with similar parameters, but some differences. The minimum sum of instance weight controls the generalisation similar to minimum sample split in GBRT. The portion of columns when constructing each tree (colsample bytree) similar to maximum features.

3.6. Performance Evaluation

Various measurements based on actual and predicted results are calculated, in order to evaluate the performance or accuracy of data-driven models. These include Coefficient of Variance (CV), Mean Bias Error (MBE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Squared Percentage error (MSPE), Mean Absolute Percentage Error (MAPE) and MAE (mean absolute error). CV is the variation of overall prediction error concerning actual mean values. MBE is used to determine the amount over/underestimation of

predictions. MSE and MSPE is a good inductor of estimation quality. MAE determines the average value of the errors in a set of forecasts and MAPE is the percentage of error per prediction. RMSE has the same unit of actual measurements. In this work, RMSE, MAE and **coefficient of determination** (R^2) are used to present the accuracy of ML models. R^2 is the percentage variance in the dependent variable explained by the independent ones. These values are calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y})^2} \tag{14}$$

$$MAE = \frac{1}{N}\sum \mid y_i - \hat{y} \mid \tag{15}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{16}$$

Here, y, $\hat{y}$ and $\bar{y}$ represent the real, estimated and average response values, respectively.

4. Selected Datasets for Case Study

Two building datasets simulated using BPS tools are utilised. First data contains 768 variations of a residential building obtained altering eight basic envelope characteristic (Tsanas & Xifara, 2012), and the second dataset includes various building type represented by 28 envelope and climate features (Rastogi, 2016). Each set and the distribution of variables are presented in this section. The prediction targets for both sets are heating and cooling loads.

4.1. Ecotect Dataset

This dataset was developed by Tsanas & Xifara (2012) and obtained from UCI machine learning repository (Xifara & Tsanas, 2012). It includes 12 residential buildings types with the same volume ($771.75m^3$) and varying envelope features, outlined in Table 1. The materials were chosen to achieve the lowest

U-values based on availability in the market (walls: 1.78 m^2K/W, floors: 0.86, roofs: 0.50 and windows: 2.26). The window-to-floor ratio is varied from 0% to 40%. The glazing distribution on each faade has 6 variants: (0) uniform, with 25% glazing on each side; (1) 55% glazing on the north faade and 15% on the rest; (2) 55% glazing on the east faade and 15% on the rest; (3) 55% glazing on the south faade and 15% on the rest; (4) 55% glazing on the west faade and 15% on the rest; and, (5) no glazing. All combinations were simulated using Ecotect with weather data from Athens, Greece, and occupancy by seven people conducting mostly sedentary activities. The ventilation was run in a mixed mode with 95% efficiency and thermostat setpoint range of 19-24$^{\circ}C$. The operating hours were set to 3 pm - 8 pm (15:00-20:00) for weekdays and 10 am - 3 pm (10:00-15:00) for weekends. The lighting level was set to 300 lx. Figure 3 illustrates the frequency distribution of features and the correlation between each pair of input and target variables is plotted as a heat map matrix in Figure 4.

Figure 3 illustrates the frequency of features as histogram graphs. The correlation between each pair of input and target variables is demonstrated using heatmap matrix in Figure 4.

4.2. EnergyPlus Dataset

This datasets consists of commercial and residential buildings and is described by Rastogi (2016). The original commercial building models were downloaded from the US Department of Energy (USDOE) commercial reference building models (DOE). The commercial buildings set includes sixteen types of buildings classified into eight overall groups based on usage. Table 2 presents the building types which are considered in the simulations and the frequency of each with unique features. For each subtype, there are three variations for envelope construction: pre-1980, post-1980, and new construction. Each usage type has the same form, area and operation schedules. The residential buildings are described by Chinazzo et al. (2015). Variation in the outputs is also introduced by considering several years of historical weather data from many

Table 1: List of features that represent the characteristics of residential buildings for prediction of energy loads

Feature	Unit	Range	Variation	Code
Inputs				
Relative compactness	-	0.62 – 0.98	12	rc
Surface area	m^2	514 – 808	12	sa
Wall area	m^2	245 – 416	7	wa
Roof area	m^2	110 – 220	4	ra
Overall height	m	3.5, 7	2	oh
Orientation	-	2 – 5	4	ori
Glazing area	m^2	0 – 0.4	4	glza
Glazing area distribution		0 – 5	6	glzd
Targets				
Heating load	KWh/m^2	6 – 43	-	heat
Cooling load	KWh/m^2	10 – 48	-	cool

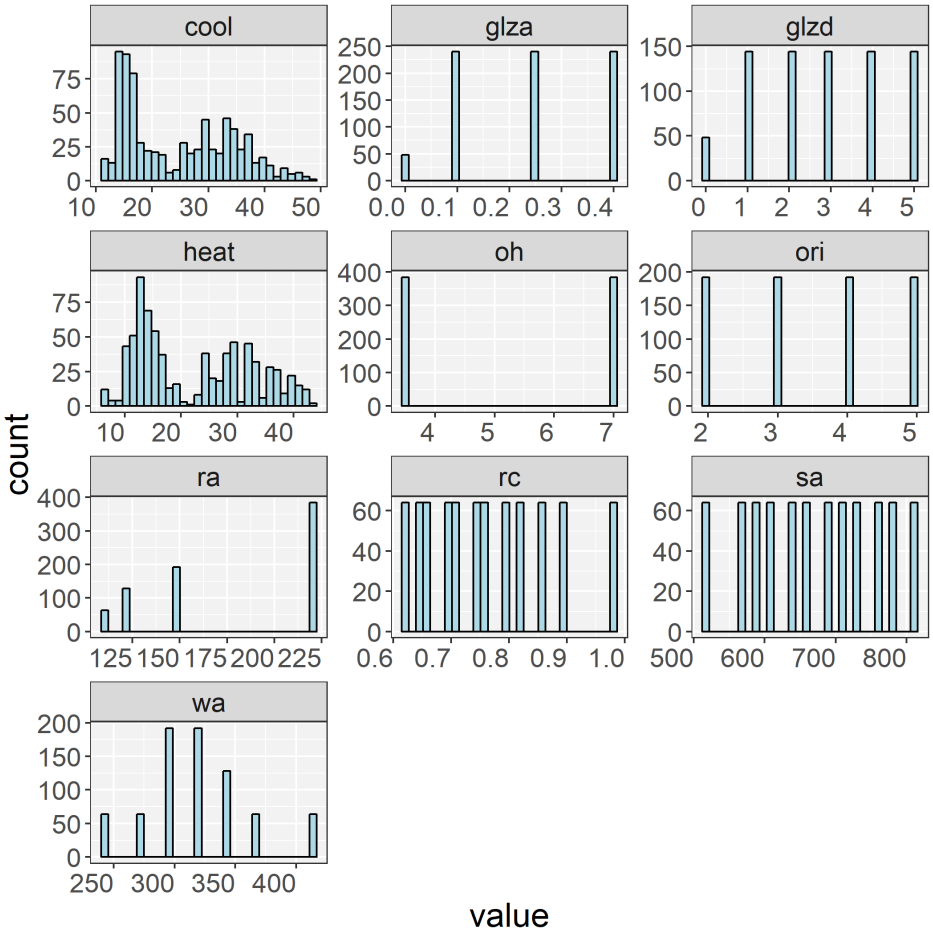


Figure 3: Distribution of features for Ecotect data.

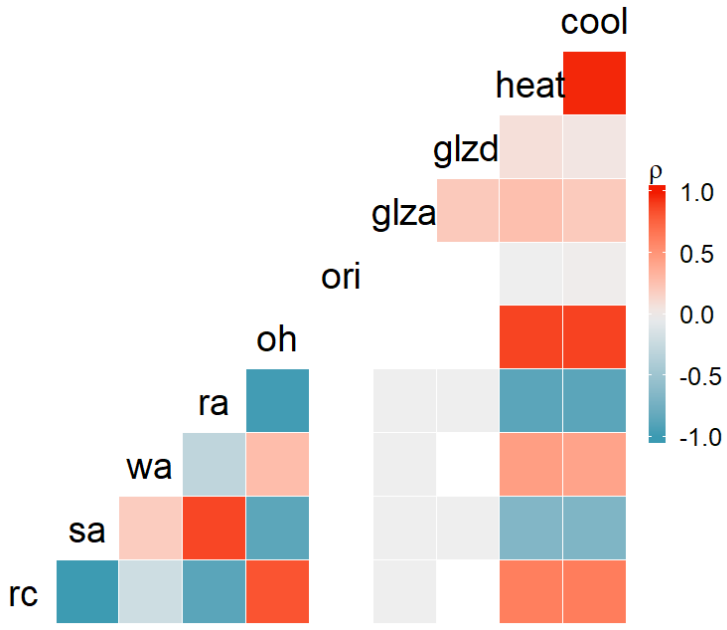


Figure 4: Ecotect data features correlation map.

climates (weather stations) and augmenting this data with synthetic weather generated for some climates (Rastogi, 2016).

We use the same regression inputs as originally proposed in (Rastogi & Andersen, 2016). They describe the feature selection as being based on correlation estimation and Principal Component Analysis (PCA). There are three kinds of input variables: climate, building, or mixed. The climate variables were extracted from one year of weather data only and are independent of the buildings simulated. The building features are related to the physical characteristics of the building envelope and independent of the climate. These inputs were chosen on the basis of impact on the heating and cooling loads and calculated from geometry, material and structure properties. The mixed parameters represent the interactions between weather and buildings. An input that does not belong to any of these categories, the internal heat gain, was also included to represent the impact of human behaviour. Figure 5 and Figure 6 illustrate the frequency of features and correlation heat-map matrix respectively.

Figure 5 illustrates the frequency of features as histogram graphs for EnergyPlus Dataset. It can be seen that the each variable is relatively distributed over the possible predefined values. The correlation heat-map matrix presented in Figure 6 shows the in dependency of different features especially building physics related ones from each other.

5. Result and Discussions

All models are implemented using Python programming language and test have been carried out on a PC with Intel Core i7-6700 3.4GHz CPU, 32GB RAM.

The stated goal of this paper is to highlight the importance of tuning nonlinear regression models (ML models) to achieve the best predictive performance for a given use case. To put our work in context, it is worth noting the results from the original studies that introduced the datasets used in this paper (Tsanas & Xifara, 2012; Rastogi et al., 2017). Tsanas & Xifara (2012) reported RMSEs of

Table 2: Frequency and size of building types in EnergyPlus data

Building Usage	Type	Area (m^2)	Volume (m^3)	No. of E+ zones	No. of samples
Health	Hospital	22,422	88,864	55	3827
	Outpatient	3,804	11,932	118	5504
Home	Mid-rise Apartment	3,135	9,553	36	37173
	Single Family	78532			
Hotel	Large	11,345	35,185	43	5504
	Small	4,014	11,622	67	5468
Office	Large	46,320	178,146	73	275345
	Medium	5503	4,982	18	19,741
	Small	511	1,559	5	5483
Restaurant	Full Service	5,502	55,035	2	3824
	Quick Service	232	708	2	5505
Retail	Stand Alone	2,294	13,993	5	5503
	Strip_Mall	2,090	10,831	10	5498
	Supermarket	45,002	900,272	6	5554
School	Primary	6,871	27,484	25	5505
	Secondary	19,592	95,216	46	5507
Warehouse	–	4,835	39,241	3	5492

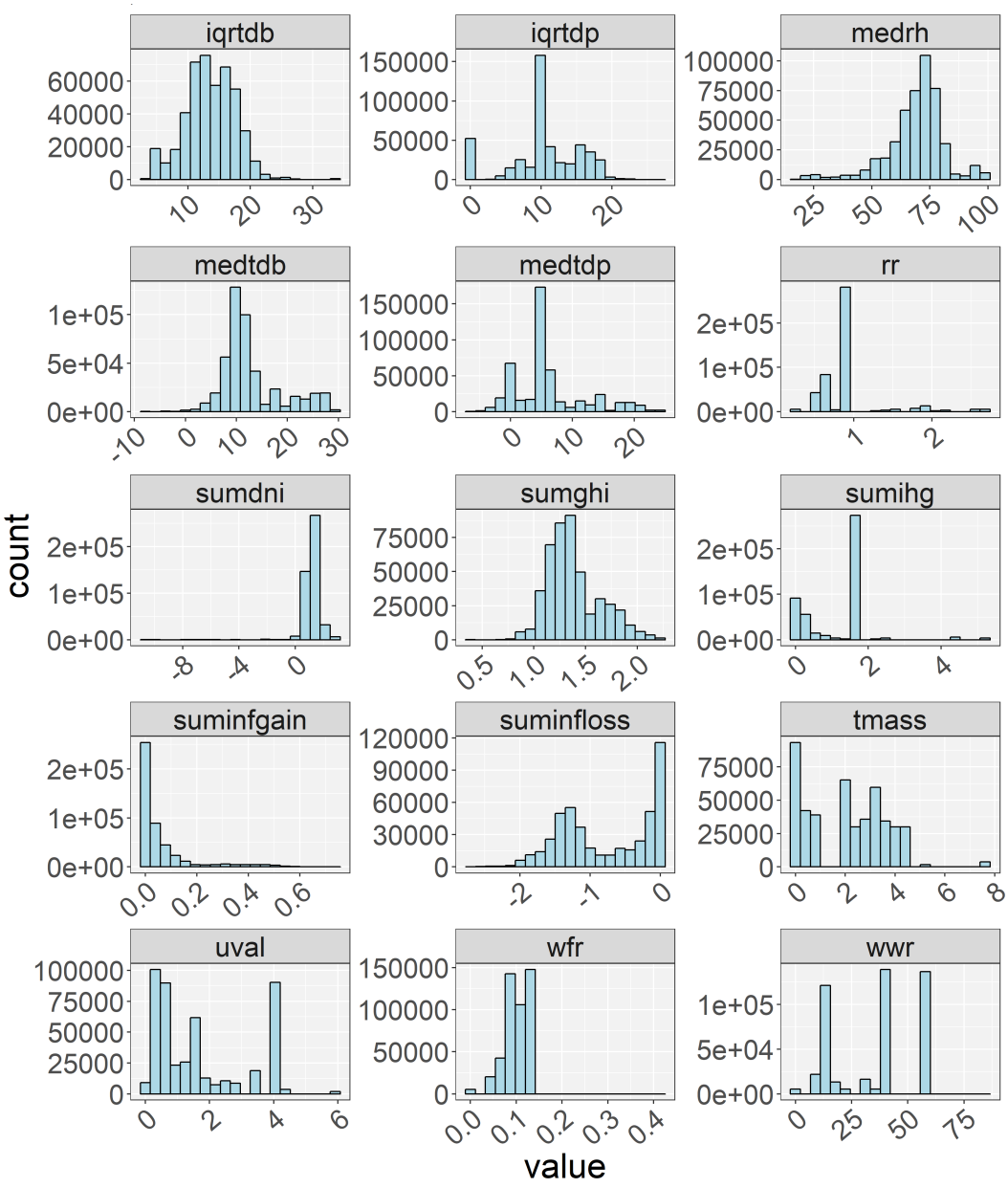


Figure 5: Distribution of features for EnergyPlus data.

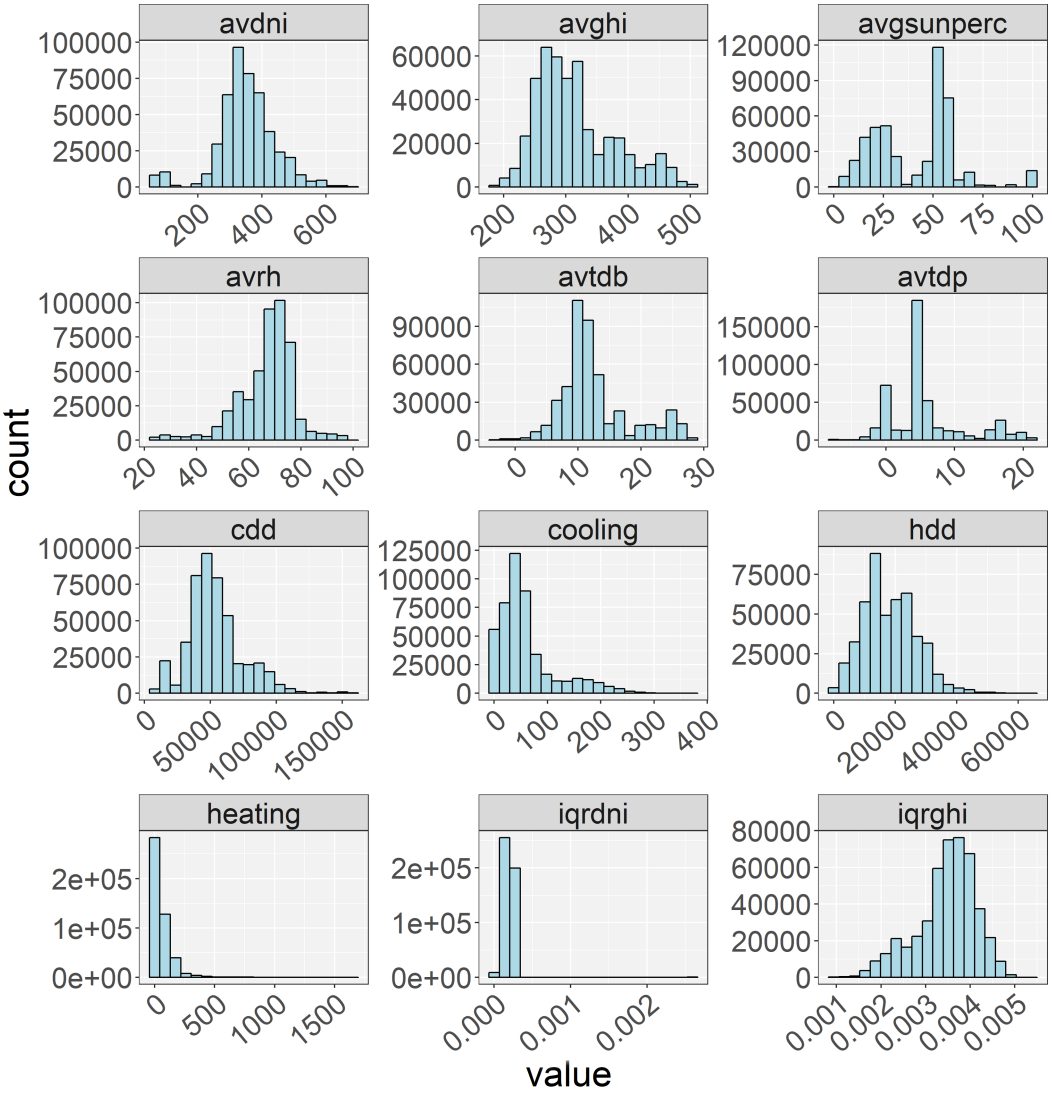


Figure 5 (Cont.): Distribution of features for EnergyPlus data.

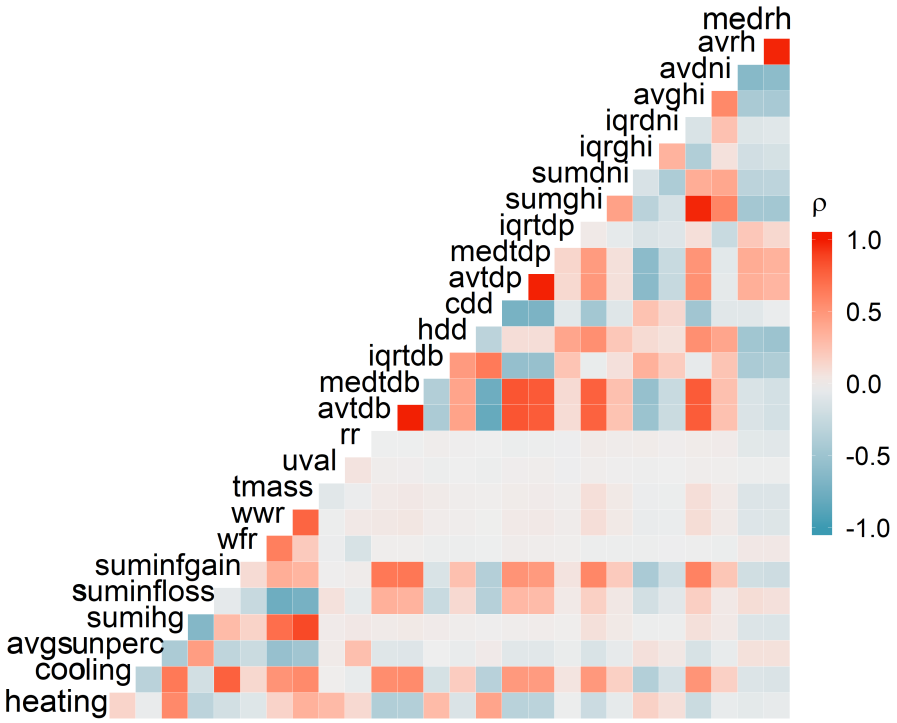


Figure 6: EnergyPlus data features correlation map.

Table 3: List of EnrgyPlus features extracted for model training

Group	QTY	Stats	Description	Range	Code	Unit
Building	U-value	Average	Average U-value of envelope	0.14–6.06	*uval*	W/m^2K
	Thermal Mass	Sum	Sum of thermal storage capacity	1e-4–7.61	*tmass*	MWh/K
	Envelope Ratios	Ratio	Ratio of window area to wall area	0.58–85.00	*wwr*	-
			Ratio of window area to floor area	0.01–0.42	wfr	-
	Massing	Ratio	Form Factor (Volume / Wall Area)	2.47–17.14	ff	-
			Roof Ratio (Roof / Wall Area)	0.31–2.73	rr	-
Mixed	Shading	Average	Average sunlit percentage of envelope	0.35–100	*avgsunperc*	%
	Infiltration	Sum	Annual sum of energy gained due to infiltration	0–0.74	*suminfgain*	GWh
			Annual sum of energy lost due to infiltration	-2.7– -1e-4	*suminfloss*	
Other		Sum	Annual sum of Internal Heat Gain	0.03–5.24	*sumIHG*	GWh

1.014 and 2.567 for heating and cooling loads, respectively. Our best RF model achieved 0.476 and 1.585 for the same variants, a roughly 40% improvement in

List of features extracted for model training (cont.)

GRP	QTY	Stats	Name	Range	Code	Unit
Climate	Degree Days	Sum	Annual sum of cooling degree days	(9.6–160)e4	*cdd*	*C-day*
			Annual sum of heating degree days	424–64878	*hdd*	
	Dry Bulb Temp (Hourly)	Avg.	Annual average of dry bulb temperature	-3.11–28.39	*avgtdb*	*C*
		Median	Median dry bulb temperature	-7.20–30	*medtdb*	
		IQR	Inter-quartile range of dry bulb Temp	3.6–34	*iqrtdb*	
	Dry Point Temp (Hourly)	Avg.	Annual average of dry point temperature	-7.41–21.43	*avgtdp*	*C*
		Median	Median dew point temperature	-6.4–24.2	*medtdp*	
		IQR	Inter-quartile range of dew point temperature	0–26.8	*iqrtdp*	
	Global Horizontal Irradiation (Hourly)	Avg.	Annual average of global horizontal irradiation	190–509	*avghi*	*MWh/m2*
		Sum	Annual sum of global horizontal irradiation	0.40–2.23	*sumghi*	
		IQR	Inter-quartile range of global horizontal irradiation	(0.84–5.2)e-3	*iqrghi*	
	Direct Normal Irradiation (Hourly)	Avg.	Annual average of direct normal irradiation	57–676	*avgdni*	*MWh/m2*
		Sum	Annual sum of direct normal irradiation	-10.34–3.15	*sumdni*	
		IQR	Inter-quartile range of direct normal irradiation	(0.38–26.3)e-4	*iqrdni*	
	Humidity (Hourly)	Avg.	Annual average of relative humidity	22–98	*avrh*	%
		Median	Median relative humidity	18–99.6	*medrh*	

accuracy in term of RMSE (kWh/m^2). Rastogi et al. (2017) report an error of 10-15 kWh/m^2 on the EnergyPlus dataset while we achieve 6-10 kWh/m^2. Tables 5 and 6 give an overview of results for the Ecotect and EnergyPlus datasets, respectively. The tables contain Coefficients of Determination (R^2), Root Mean Square Errors (RMSE), Mean Absolute Errors (MAE), fit time, test time, and number of parameters for the best combination of hyper-parameters; the average fitting time of all tested models; and the total number of iterations for comparison of time complexity. Here, the test time is the average of predictions of all folds (192 data points per fold for Ecotect and 1,000 for Energy Plus). For EnergyPlus data, GP is excluded from the comparison because the training time is extremely high for large datasets. Most ML models are capable of forecasting multiple outputs at the same time. However, we tuned all models separately for heating and cooling loads. None of the techniques obtained the best accuracies for both target values using the same combination of hyper-parameters. This inconsistency indicates that the importance of input variables as well as the corresponding weights are different. Hence, two independent models are required, rather than training a single model.

Though the datasets are drawn from different simulators, similarities in the performance of the models do emerge. The lowest RMSE for both heating and cooling loads is achieved by XGBoost, followed by GBRT and RF. These models are all based on decision trees, but unlike RF the other two do not build independent trees. Hence, they train models slightly faster than RF. Considering prediction time in addition to accuracy, however, GBRT is slightly faster than XGBoost but has comparable accuracy. The NN models tend to have the fastest prediction times, which might make them more appropriate for applications requiring very large numbers of simulation estimates. For example, optimising many building parameters, each with several possible choices, under a sample of uncertain operating conditions, such as the problem described in (Rastogi et al., 2017). We find that GP is the slowest and least accurate model. This is partially due to the challenge of using large datasets with GP regression; since the time complexity of GP is $O(N^3)$ (where N is the number of data points

Table 5: Result of tuning ML model for Ecotect simulated dataset.

	SVM		RF		NN		GP		GBRT		XGBoost	
	Heat	Cool	Heat	Cool	Heat	Cool	Heat	Cool	Heat	Cool	Heat	Cool
R^2	0.996	0.972	0.998	0.973	0.997	0968	0.981	0.944	0.999	0.995	0.999	0.998
RMSE	0.475	1.622	0.476	1.585	0.491	1.711	1.381	2.279	0.366	0.677	0.300	0.401
MAE	0.654	1.082	0.332	0.98	0.369	1.120	0.852	1.579	0.254	0.486	0.189	0.294
Fit time (s)	2.19	21.09	0.75	0.73	0.106	0.131	18.86	24.63	0.499	0.655	0.326	0.585
Mean fit time (s)	323.77	177.28	0.72	0.75	1.23	0.99	17.00	18.65	0.20	0.19	0.28	0.28
Test time	0.005	0.005	0.090	0.104	0.001	0.001	0.019	0.018	0.021	0.031	0.045	0.107
Number of parameters	2		3		7		3		7		6	
Total iteration	21		36		3240		30		3456		2160	

Table 6: Result of tuning ML model for 5000 records of EnergyPlus dataset.

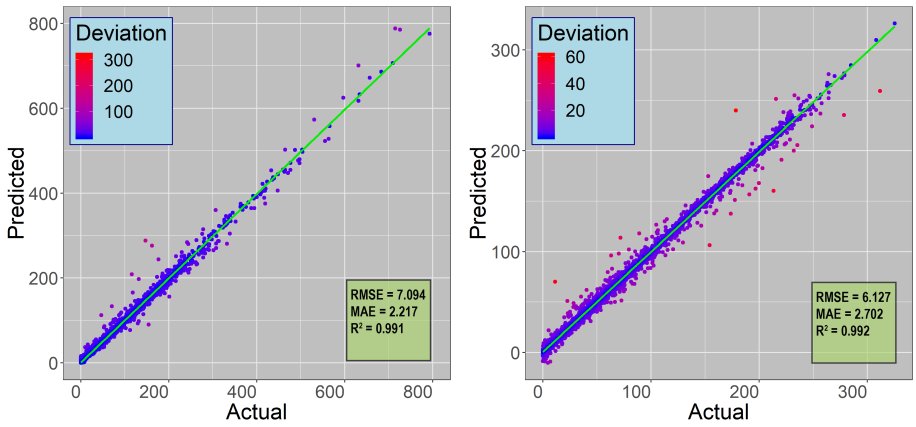
	SVM		RF		NN		GBRT		XGBoost	
	Heat	Cool	Heat	Cool	Heat	Cool	Heat	Cool	Heat	Cool
R^2	0.965	0.973	0.973	0.968	0.966	0969	0.980	0.986	0.982	0.986
RMSE	14.318	8.763	12.720	9.400	14.068	9.376	10.721	6.296	10.386	6.270
MAE	5.622	3.465	5.057	4.841	7.472	4.932	4.400	3.365	4.130	3.143
Fit time (s)	177.66	406.31	6.35	34.873	126.29	10.88	6.363	1.789	4.897	4.871
Mean fit time (s)	1641.91	1197.16	17.6	19.54	21.32	17.19	4.85	4.92	4.61	4.55
Test time	0.483	0.507	0.333	0.595	0.008	0.010	0.244	0.078	0.228	0.219
Number of parameters	2		3		7		3		7	
Total iteration	21		36		3240		3456		2160	

used for training/fitting), the training speed is not comparable with other ML models and inversion of matrices of size $\{N, N\}$ is unfeasible for large N. Thus, studies using GP have used small datasets, usually less than a few thousand (Heo et al., 2012; Zhang et al., 2013; Noh & Rajagopal, 2013; Rastogi et al., 2017; Burkhart et al., 2014; Zhang et al., 2015). However, since GP regression allows for the automatic estimation of prediction uncertainty, it is useful in some cases. An example is the estimation of summary statistics, where it is more informative to know the uncertainty of, e.g., annual heating and cooling loads, rather than just a point estimate. Although all models predict the energy loads with high accuracy, the use case should determine the most appropriate model. For example, increasing the number of records (size of training data), the fitting and forecasting time of SVM rises significantly. The training size of NN is slightly increased as well, but it is still the fastest predictor by a considerable margin (10-20 times faster). GBRT and its variant XGBoost achieve the best RMSE. However, the increased accuracy and sophistication of models like NN and XGBoost comes with the penalty of requiring very large training datasets (e.g., the 25,000 simulations used here). This could be an issue where a model has to trained on the fly, i.e., where simulating 25,000 distinct cases to train an accurate model is prohibitively expensive. As expected, using more data to fit a model increases the predictive accuracy of all models, such that complex models with more parameters lose out to simpler models that have seen more data, provided the simpler models can use the additional data available. In summary, where sufficient training data is available and the testing or use cases are not too dissimilar from the training data, the use of models such as GBRT and NN improves accuracy. Where training data is harder to generate, or a model must be trained on the fly with a small dataset, techniques such as GP provide adequate predictions.

5.1. Performance of the Best Model

We now discuss the performance characteristics of the best models for each dataset. The results are illustrated using two kinds of plots: predicted (esti-

mated) loads ($\hat{y}$) against loads from the simulator (y), and the distribution of errors between simulated-predicted pairs ($\hat{y} - y$). Figures 9 and 7 show the values predicted by tuned GBRT models against their corresponding simulated heating and cooling loads for the Ecotect and EnergyPlus datasets, respectively. The error distributions of these estimations are given in Figure 10 and 8.

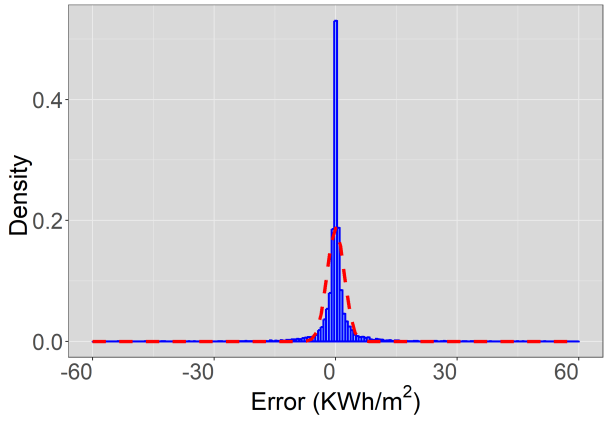


(a) (b)

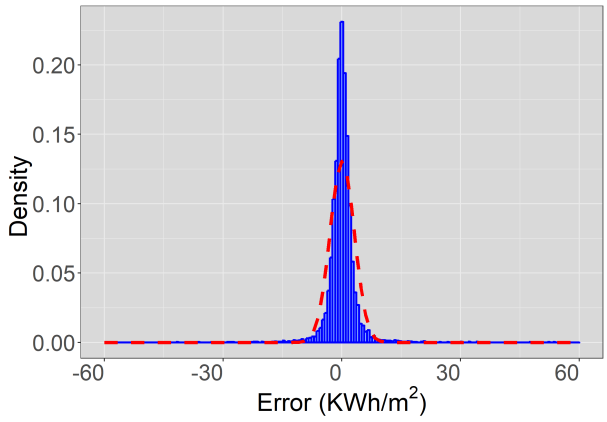
Figure 7: Actual and predicted (a) heating and (b) cooling loads of EnergyPlus dataset.

5.2. Effect of Increasing Size of Training Data

Given that using large datasets for training seems to improve the predictive accuracy of all models, we investigated the effect of increasing the size of the training dataset on accuracy. Figure 11 shows RMSE versus size of training dataset for the GBRT model. A 10 fold cross-validation is used to obtain the worst, best and mean RMSE over all folds. Mean training time is also displayed as the top axis to show computational cost. Although the best result is obtained by the highest number of samples tested, 25,000 is enough to build a reliable model considering the fitting time and error gap. At this point, the mean RMSE is equal to 7.770 kWh/m^2 and time required to fit the model is 66.02 seconds. On the other hand, using 400,000 samples and fitting over 2600 seconds, mean RMSE only goes down to 2.338 kWh/m^2 (4% of average heating loads).

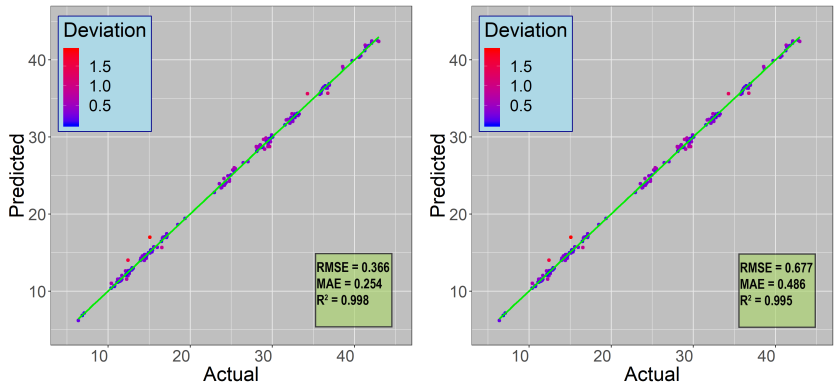


(a)



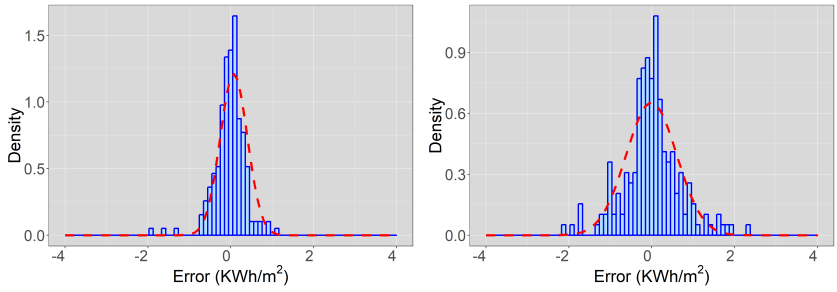
(b)

Figure 8: Error distribution of (a) heating and (b) cooling loads prediction for EnergyPlus dataset.



(a) (b)

Figure 9: Actual and predicted (a) heating and (b) cooling loads of Ecotect dataset using GBRT model.



(a) (b)

Figure 10: Error distribution of (a) heating and (b) cooling loads prediction for Ecotect dataset. The red dashed line is for a theoretical normal PDF with the same parameters.

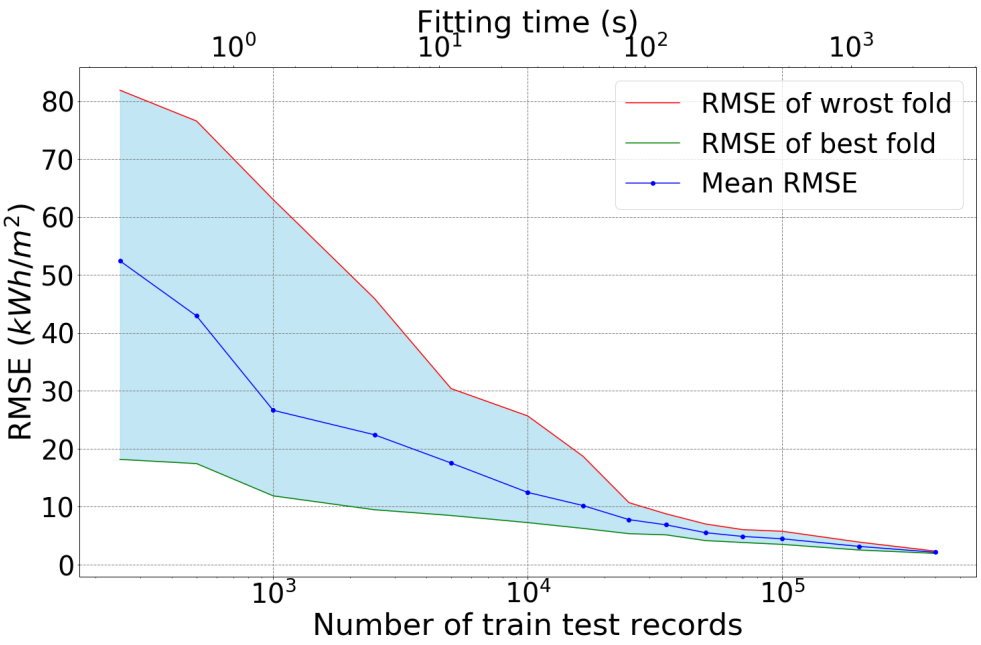


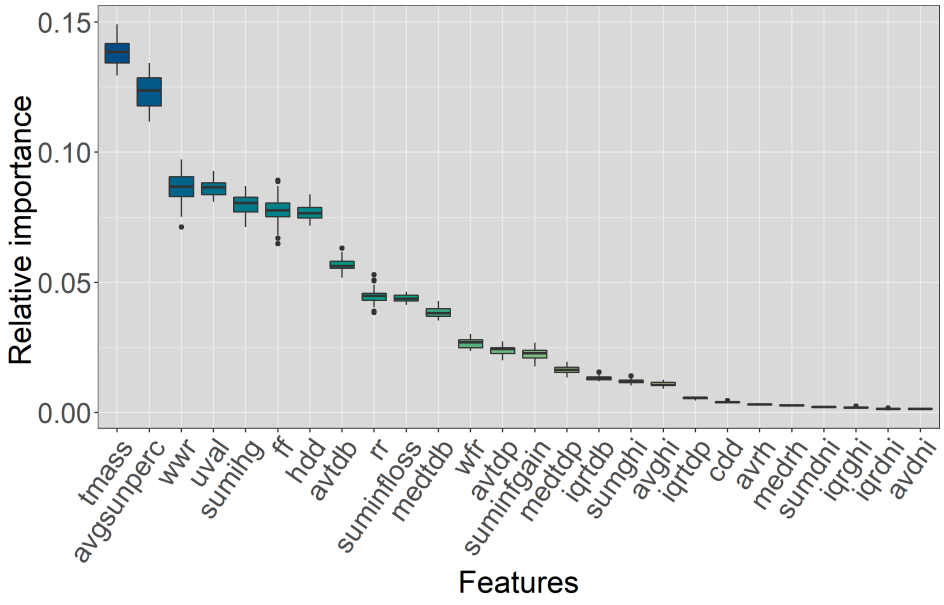
Figure 11: RMSE for heating load against number of total number of samples used for training.

5.3. Feature Importance and Selection

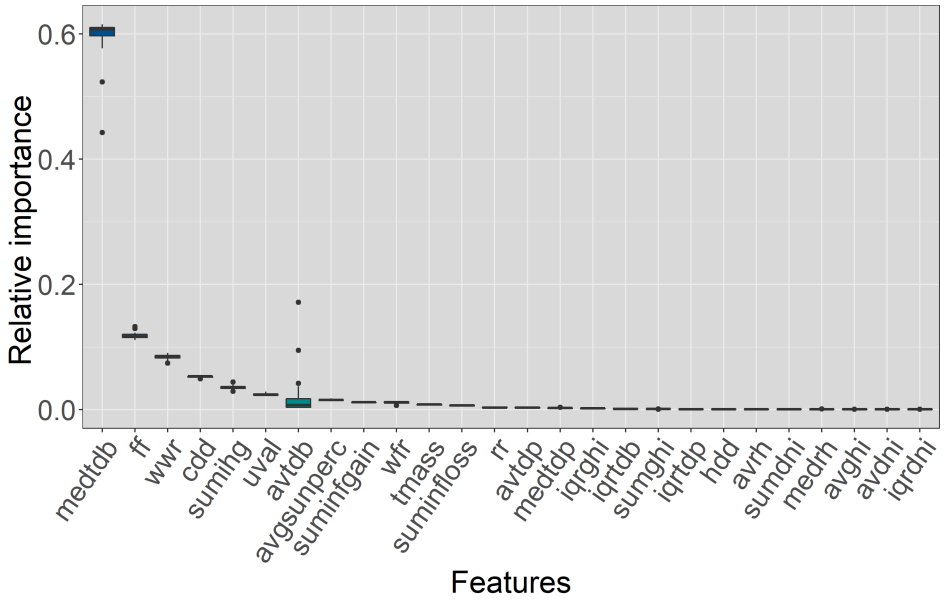
To emphasise the importance of features in predicting different loads, we present a sensitivity analysis using two approaches. First, we present the feature importance calculated by the RF models. RF creates many decision trees and the amount of weighted variance explained by each feature can be calculated for each tree. For a forest, the variance explained by each feature can be averaged and the features ranked according to this measure. Here, we trained 30 RF models using 100,000 randomly selected samples to obtain an empirical distribution of feature importance, shown in Figure 12.

As the best model (GBRT) doesn't provide the possibility of analysing sensitivity to the input variables in the same way, we used a global variance-based method called the Sobol method Sobol (2001); Saltelli (2002). Unlike RF, GBRT does not generate unique trees. Rather, each trees is correlated to the last. To facilitate a comparison, we fitted 30 different models and used them to evaluate the 150,000 samples generated by the algorithm. The Sobol first-order indices of features is illustrated in Figure 13. We see that the importance of features to this method is less stable in GBRT than RF. However, since it is calculated directly from the original data, it is more representative of the features of the dataset itself.

For a final test, we examined the effect of dropping variables that the model does not deem to be important. Based on the results of the Sobol comparison, we identified the following features to drop: 'avrh', 'avdni', 'iqrdni', 'iqrghi', 'medrh', 'sumdni' for both loads, and 'avghi', 'sumghi' for cooling only. All of the dropped variables are climate-related, which implies that there may be too many variables used to explain variance due to climate. The GBRT with fewer features was also trained and tested over 10 folds with 25,000 random samples. The results of training a model with a reduced feature set is compared with using the full set of features in Table 7. We see that removing features to which the model is apparently insensitive does not negatively affect the model performance. However, the time complexity of training model is reduced due to a reduction in the size of the dataset. Given that this result applies only to this

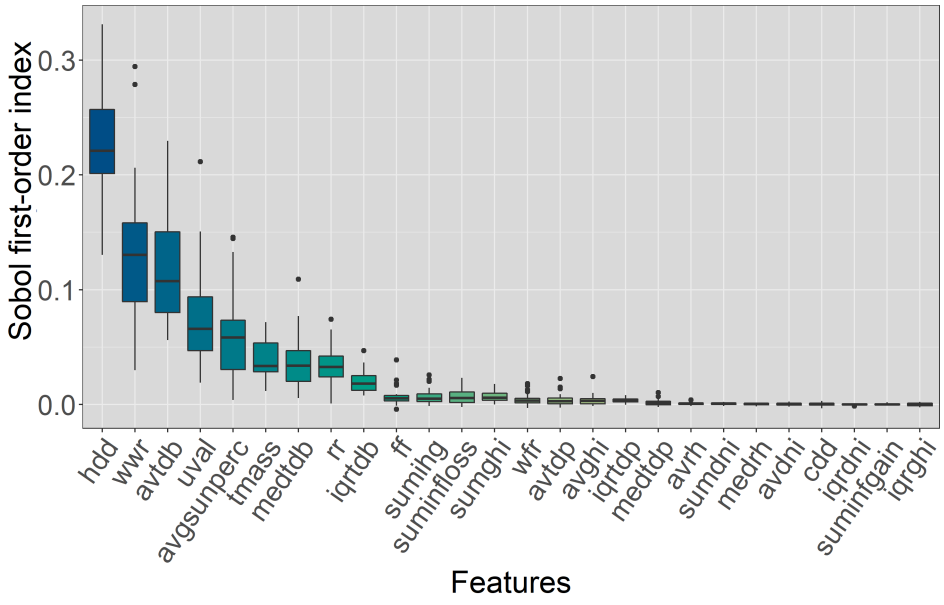


(a)

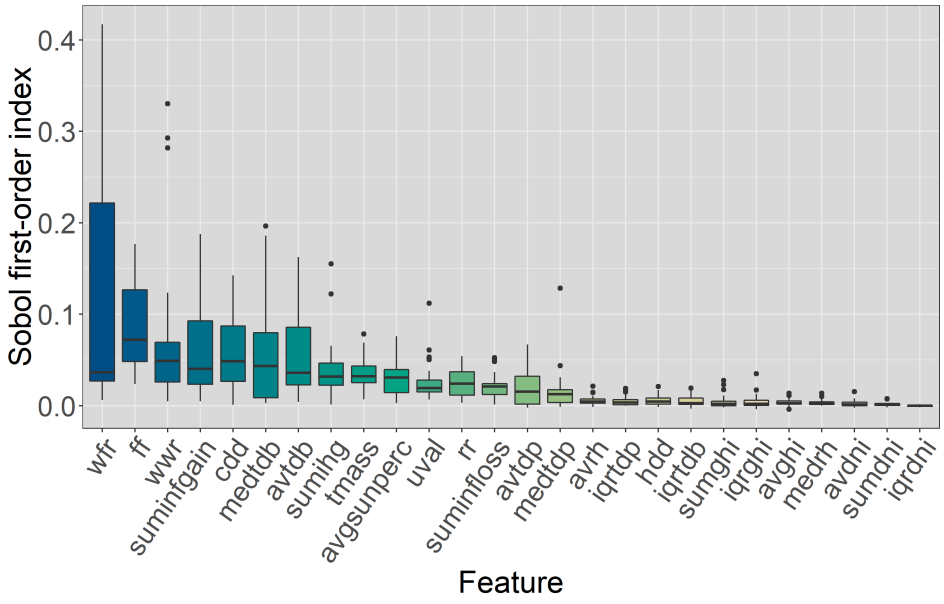


(b)

Figure 12: Importance of features for (a) heating and (b) cooling loads prediction using RF model.



(a)



(b)

Figure 13: Sobol first-order indices of features in predicting (a) heating and (b) cooling loads using best ML model.

dataset and cannot be generalised to all buildings or EnergyPlus simulations, this is not a repeatable result unless we are confident that the dataset used for training represents the use case or problem completely.

Table 7: Performance comparison of ML models with full and reduced feature sets determined by sensitivity analysis.

	Heating Load		Cooling Load	
	All inputs	Selected inputs	All inputs	Selected inputs
RMSE	7.871	7.648	4.455	4.384
MAE	2.127	2.085	2.314	2.310
R^2	0.991	0.991	0.993	9.993
Fit time (s)	61.621	48.420	9.387	7.700
Test time (s)	0.642	0.622	0.151	0.145

6. Conclusion

The research presented in this paper addresses the gap in using ML methods for estimating building energy loads through a comprehensive study of common ML models fitting over energy simulation data. As became evident in the reviewed literature, despite the wide usage of MLs in this field, a conclusion on selecting the right model for the energy prediction was not possible. The main reason is that most of the research works have focused on the first eminent part of statistical modelling which is features selection. This paper discussed the importance of ML model optimisation in providing a fair comparison of different methods in term of accuracy, the simplicity of tuning and training and response times of model. This study optimised the hyper-parameters of each model for both heating and cooling loads to obtain the best precision. It was also indicated that when there are two energy indices as cooling and heating loads to be estimated by model, it is desired to optimise and train separate

machines. To that end, the role of ML model in recognising most impacting factor in prediction of building loads. The other key outcome of this research is a set of recommendations for the quick selection of ML model based on the data and usage.

The results indicated that the standard and advanced GBRTs provide the most accurate predictions, considering the RMSE value. However, when the data was simple (in term of input variables and size), SVM was proven to be the best choice because of simplicity and the speed of calculations. The results also ascertained that for complex data sets, multi-layer NNs are more appropriate when there is a massive demand for ever-more energy simulations. In this case, NN was proven to be capable of estimating incredibly faster than other MLs methods. It should be noted that NN is complicated, and requires an expert to particularly tune it for each studied case; otherwise, NNs could fail quickly.

Comparison of tuned models with previous studies highlighted the importance of determining the hyper-parameters for each data set, and the fact that this can become more crucial by increasing the size and intricacy of the examinations set. By fitting individual models for heating and cooling loads, it was shown that one assorted set of model parameters could not accurately estimate both values. Therefore, unlike previous studies, it is recommended by this study to train models for each energy load independently. The other approach would be the implementation of a specific sorting algorithm to find balanced values. As results signified, it is suggested to attain a higher accuracy feeding the machines with more number of instances is essential. It might not be a solution for measured historical data; however further simulation using various values of inputs could be aggregated during the design stage prior to optimising the building. Another identified critical factor was that the features must be thoroughly selected/created for representing building characteristics and needs should be appropriately investigated before developing models.

The findings of this study concurred with the seminal literature by demonstrating the fact that MLs techniques are overtly superior over the conventional statistical and engineering methods in building energy calculation. This study

also revealed the further power of those ML methods and newly developed ones when they thoroughly optimised. There are several ready to use software packages (e.g. Matlab) providing various ML models with few parameters to modify. Nevertheless, it is advisable to use simpler models like SVM or RF with an advanced programming language, such as Python and R.

Finally, the most important features are recognised using sensitivity analysis methods, and the investigation of the model with reduced dimension revealed that even though the computational cost of building model is reduced, the performance didn't alter. This analysis demonstrated the capability of MLs in eliminating inessential input parameters, while most statistical methods are susceptible to these type of features.

The methods discussed in this work proved the efficiency of ML models in predicting building energy loads as well as performance. The fast and accurate calculation of those values paves pathways for more informed and productive design decisions for built environments. Furthermore, along with the optimisation algorithms, ML seems as a promising solution for efficacious retrofit planning of complex buildings, where engineers are not capable of massive calculations.

Appendix A. Detailed Results for Tuning ML Models

The detail of tuning each ML model is presented in this section. Some models have several parameters, so the brute force search includes thousands of train-test models. Therefore, it is not possible to present the list of all results in this paper. However,Tables A.8 to A.13 demonstrates the parameters for the best models predicting energy loads of both datasets. In each table the best model is highlighted with light blue colour.

In order to reduce the time complexity of tuning ANN model, the number of epochs was fixed at 500 and the other parameters were optimised. Then the optimal number of propagations was separately obtained using the best parameters. As shown in the Figures A.14 (a) and (b)

Table A.8: Detail of optimising SVM for both datasets.

EPlus Data		Ecotect Data		SVM Parameters	
Heat RMSE	Cool RMSE	Heat RMSE	Cool RMSE	C	Gamma
14.318	9.785	0.677	1.622	10,000	1
18.988	9.774	0.654	1.667	1000	1
15.720	9.261	0.660	1.756	1,000,000	0.1
15.313	10.302	0.978	1.842	1,000,000	1
21.626	8.763	0.815	2.048	100,000	0.1
31.415	9.452	2.108	2.636	10,000	0.1
43.719	17.833	2.627	3.365	10,000	0.01
60.974	31.658	3.304	3.886	1	0.1
60.974	31.658	3.304	6.550	1	0.01

Table A.9: Detail of optimising RF for both datasets.

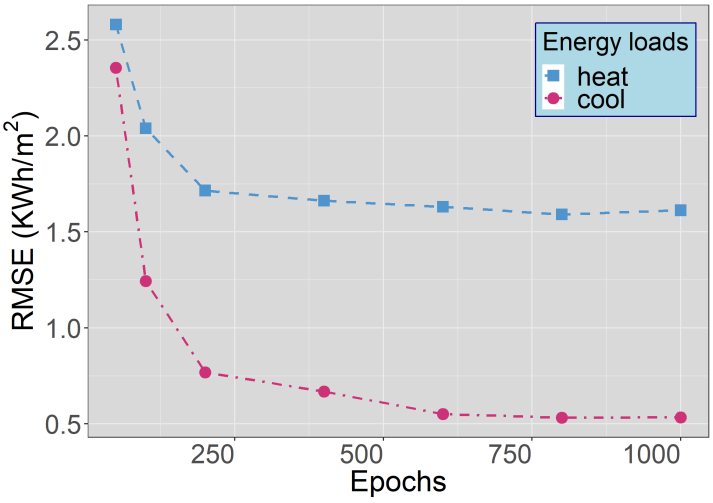
EPlus Data		Ecotect Data		RF Parameters		
Heat RMSE	Cool RMSE	Heat RMSE	Cool RMSE	Boot-strap	Max features	No. of es-timators
12.873	9.894	0.568	1.585	False	sqrt	600
12.720	9.693	0.576	1.605	False	sqrt	400
14.556	10.734	0.604	1.612	True	sqrt	200
13.334	10.214	0.502	1.658	False	log2	1000
14.551	9.691	0.476	1.683	True	auto	600
14.584	9.600	0.478	1.691	True	auto	800
24.189	13.727	0.536	1.814	False	auto	1000
14.199	10.995	0.616	1.604	True	sqrt	400

Table A.10: Detail of optimising GBRT for both datasets.

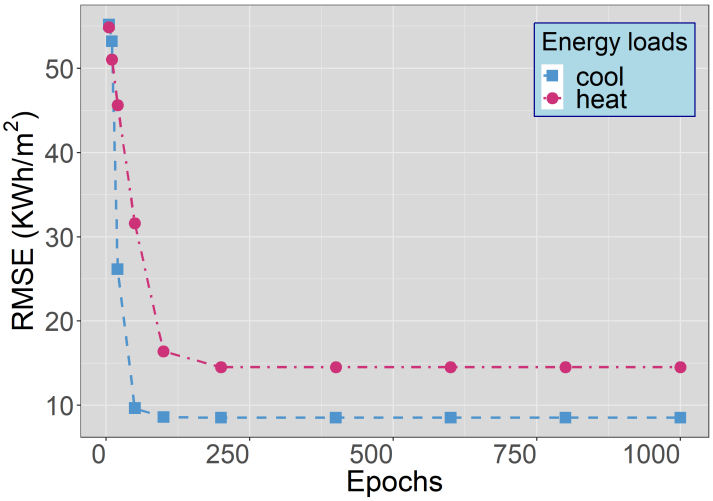
EPlus Data		Ecotect Data		GBRT Parameters						
Heat RMSE	Cool RMSE	Heat RMSE	Cool RMSE	Learning rate	Max depth	max features	Min sample leaf	Min sample split	No. estimator	Sub sample
13.157	7.402	0.388	0.677	0.15	8	None	1	100	1500	1
11.534	6.667	0.366	0.893	0.15	3	None	1	2	1500	1
12.914	6.297	0.399	1.033	0.25	3	sqrt	1	2	1500	1
10.721	8.855	0.514	1.261	0.1	8	sqrt	1	2	2000	1
13.257	9.608	0.523	1.578	0.01	13	sqrt	1	2	1250	0.9
17.843	8.654	3.77	3.895	0.01	8	sqrt	200	100	500	1
22.775	9.856	8.732	8.524	0.001	13	sqrt	200	100	1250	0.7
26.221	15.625	10.054	9.725	0.25	3	sqrt	1000	100	1750	1
35.518	21.196	10.003	9.726	0.1	8	sqrt	1000	1000	1000	0.9
46.991	24.279	9.996	9.720	0.1	3	sqrt	1000	2	1000	0.8
49.515	24.542	9.996	9.722	0.1	3	None	1000	100	1500	0.7
21.578	11.642	10.422	9.746	0.15	13	None	500	1000	1000	0.8

Table A.11: Detail of optimising XGBoost for both datasets.

EPlus Data		Ecotect Data		XGBoost Parameters					
Heat RMSE	Cool RMSE	Heat RMSE	Cool RMSE	Portion of features	Learning rate	Max depth	Min child weight	No. estimator	Sub sample
10.909	8.145	0.303	0.401	0.6	0.1	8	1	1750	0.9
11.616	9.002	0.300	0.452	0.6	0.1	13	1	750	0.7
12.273	8.048	0.323	0.573	0.38	0.1	8	1	1250	0.8
12.436	8.18	0.329	0.804	0.6	0.01	13	1	1750	1
11.302	6.270	0.413	1.131	0.6	0.1	3	1	2000	0.7
10.387	6.382	0.409	1.115	0.6	0.1	3	3	2000	0.9
13.738	9.982	0.306	0.443	0.5	0.1	13	1	1750	0.8
16.030	12.706	0.337	0.557	0.4	0.5	13	3	1000	0.7
20.003	10.434	0.304	0.433	0.1	0.1	8	1	10000	0.8
23.821	14.138	0.343	0.567	0.1	0.25	13	3	1000	0.7
26.266	16.581	0.365	0.578	0.1	0.5	13	3	1500	0.8
57.263	35.554	9.459	10.412	0.6	0.01	3	1	200	0.7



(a)



(b)

Figure A.14: RMSE of ANN model predicting energy loads for (a) EPlus and (b) Ecotects datasets against number of epochs.

Table A.12: Detail of optimising GP for both datasets.

Ecotect Data		GP Parameters		
Heat RMSE	Cool RMSE	Alpha	Kernel	No. restarts
1.382	2.279	1e-08	Mattern	2
1.381	2.383	1e-12	RBF	4
8.472	2.332	1e-8	RBF	2
8.471	2.333	1e-10	RBF	0
1.383	3.138	1e-4	Mattern	0
4.440	4.238	1e-6	RBF	4

References

Ascione, F., Bianco, N., De Stasio, C., Mauro, G. M., & Vanoli, G. P. (2017). Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. *Energy*, *118*, 999–1017. URL: `http://dx.doi.org/10.1016/j.energy.2016.10.126`. doi:`10.1016/j.energy.2016.10.126`.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Bukkapatnam, S. T., & Cheng, C. (2010). Forecasting the evolution of nonlinear and nonstationary systems using recurrence-based local Gaussian process models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *82*, 56206. doi:`https://doi.org/10.1103/PhysRevE.82.056206`.

Burkhart, M. C., Heo, Y., & Zavala, V. M. (2014). Measurement and verification of building systems under uncertain data: A Gaussian process modeling approach. *Energy and Buildings*, *75*, 189–198. URL: `http://dx.doi.org/10.1016/j.enbuild.2014.01.048`. doi:`10.1016/j.enbuild.2014.01.048`.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge*

Table A.13: Detail of optimising ANN for both datasets.

EPlus Data		Ecotect Data		ANN Parameters				
Heat RMSE	Cool RMSE	Heat RMSE	Cool RMSE	Activa-tion	Batch size	No. hidden layers	No. neurons	Optimiser
30.456	12.103	0.492	1.829	Logistic	10	1	2N	lbfgs
30.675	15.847	0.781	1.711	Tanh	10	1	2N	lbfgs
14.631	9.973	0.884	3.389	Tanh	1	2	N, 2N	adam
17.337	9.376	1.052	3.308	Tanh	10	3	3N, 2N, N	lbfgs
14.068	9.754	2.616	3.388	Tanh	1	1	2N	adam
41.157	18.079	4.476	6.645	Relu	1	2	N, 2N	adam
27.573	13.994	10.005	9.73	Logistic	10	3	3N, 2N, N	adam
51.323	26.502	14.453	15.968	Relu	auto	1	N	adam
49.118	25.223	6.401	7.010	Identity	1	2	N, N/2	adam
80.052	57.114	23.506	25.482	Logistic	10	2	3N, N/2	adam

discovery and data mining (pp. 785–794). ACM. doi:https://doi.org/10.1145/2939672.2939785.

Chen, Y., & Tan, H. (2017). Short-term prediction of electric demand in building sector via hybrid support vector regression. *Applied Energy*, *204*, 1363–1374. doi:https://doi.org/10.1016/j.apenergy.2017.03.070.

Chinazzo, G., Rastogi, P., & Andersen, M. (2015). Assessing robustness regarding weather uncertainties for energy-efficiency-driven building refurbishments. *Energy Procedia*, *78*, 931–936.

Deb, C., Eang, L. S., Yang, J., & Santamouris, M. (2016). Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. *Energy and Buildings*, *121*, 284–297. URL: http://dx.doi.org/10.1016/j.enbuild.2015.12.050. doi:10.1016/j.enbuild.2015.12.050.

DOE (). Commercial Reference Buildings. URL: https://www.energy.gov/eere/buildings/commercial-reference-buildings.

Dombayci, Ö. A. (2010). The prediction of heating energy consumption in a model house by using artificial neural networks in Denizli-Turkey. *Advances in Engineering Software*, *41*, 141–147. doi:https://doi.org/10.1016/j.advengsoft.2009.09.012.

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, *37*, 545–553. doi:10.1016/j.enbuild.2004.09.009.

Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, *49*, 591–603. URL: http://dx.doi.org/10.1016/j.enbuild.2012.03.010. doi:10.1016/j.enbuild.2012.03.010.

Ghiassi, M., Saidane, H., & Zimbra, D. (2005). A dynamic artificial neural network model for forecasting time series events. *International Journal of Fore-*

casting, *21*, 341–362. doi:https://doi.org/10.1016/j.ijforecast.2004.10.008.

Grosicki, E., Abed-Meraim, K., & Hua, Y. (2005). A weighted linear prediction method for near-field source localization. *IEEE Transactions on Signal Processing*, *53*, 3651–3660. doi:https://doi.org/10.1109/TSP.2005.855100.

Harpham, C., & Dawson, C. W. (2006). The effect of different basis functions on a radial basis function network for time series prediction: a comparative study. *Neurocomputing*, *69*, 2161–2170. doi:https://doi.org/10.1016/j.neucom.2005.07.010.

Heo, Y., Choudhary, R., & Augenbroe, G. A. (2012). Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings*, *47*, 550–560. URL: http://dx.doi.org/10.1016/j.enbuild.2011.12.029. doi:10.1016/j.enbuild.2011.12.029.

Heo, Y., & Zavala, V. M. (2012). Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings*, *53*, 7–18. URL: http://dx.doi.org/10.1016/j.enbuild.2012.06.024. doi:10.1016/j.enbuild.2012.06.024.

Hong, S. M., Paterson, G., Burman, E., Steadman, P., & Mumovic, D. (2014a). A comparative study of benchmarking approaches for non-domestic buildings: Part 1 Top-down approach. *International Journal of Sustainable Built Environment*, *2*, 119–130. URL: http://dx.doi.org/10.1016/j.ijsbe.2014.04.001. doi:10.1016/j.ijsbe.2014.12.001.

Hong, S.-M., Paterson, G., Mumovic, D., & Steadman, P. (2014b). Improved benchmarking comparability for energy consumption in schools. *Building Research & Information*, *42*, 47–61. URL: http://www.tandfonline.com/doi/abs/10.1080/09613218.2013.814746. doi:10.1080/09613218.2013.814746.

Hou, Z., & Lian, Z. (2009). An application of support vector machines in cooling load prediction. In *Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on* (pp. 1–4). IEEE.

Jain, R. K., Smith, K. M., Culligan, P. J., & Taylor, J. E. (2014). Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, *123*, 168–178. URL: http://dx.doi.org/10.1016/j.apenergy.2014.02.057. doi:10.1016/j.apenergy.2014.02.057.

Kalogirou, S., Florides, G., Neocleous, C., & Schizas, C. (2001). Estimation of Daily Heating and Cooling Loads Using Artificial Neural Networks. In *Proceedings of CLIMA 2000 International Conference* September (pp. 15–18). Naples.

Kalogirou, S., Neocleous, C., & Schizas, C. (1997). Building Heating Load Estimation Using Artificial Neural Networks. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques* (pp. 1–8). volume 8. URL: http://www.inive.org/members{_}area/medias/pdf/Inive/clima2000/1997/P159.pdf.

Khayatian, F., Sarto, L., & Dall'O', G. (2016). Application of neural networks for evaluating energy performance certificates of residential buildings. *Energy and Buildings*, *125*, 45–54. URL: http://dx.doi.org/10.1016/j.enbuild.2016.04.067. doi:10.1016/j.enbuild.2016.04.067.

Leung, H., Lo, T., Member, S., & Wang, S. (2001). Prediction of Noisy Chaotic Time Series Using an Optimal Radial Basis Function Neural Network. *IEEE Transactions on Neural Networks*, *12*, 1163–1172. doi:https://doi.org/10.1109/72.950144.

Li, K., Hu, C., Liu, G., & Xue, W. (2015). Building's electricity consumption prediction using optimized artificial neural networks and principal component

analysis. *Energy and Buildings*, *108*, 106–113. URL: http://dx.doi.org/10.1016/j.enbuild.2015.09.002. doi:10.1016/j.enbuild.2015.09.002.

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. *Energy Conversion and Management*, *50*, 90–96. URL: http://dx.doi.org/10.1016/j.enconman.2008.08.033. doi:10.1016/j.enconman.2008.08.033.

Li, Q., Ren, P., & Meng, Q. (2010). Prediction model of annual energy consumption of residential buildings. In *Advances in Energy Engineering (ICAEE), 2010 International Conference on* (pp. 223–226). IEEE.

Li, Z., Han, Y., & Xu, P. (2014). Methods for benchmarking building energy consumption against its past or intended performance: An overview. URL: http://www.sciencedirect.com/science/article/pii/S0306261914002505. doi:10.1016/j.apenergy.2014.03.020.

Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings*, *92*, 322–330. doi:10.1016/j.enbuild.2015.02.007.

Mena, R., Rodríguez, F., Castilla, M., & Arahal, M. R. (2014). A prediction model based on neural networks for the energy consumption of a bioclimatic building. *Energy and Buildings*, *82*, 142–155. URL: http://dx.doi.org/10.1016/j.enbuild.2014.06.052. doi:10.1016/j.enbuild.2014.06.052.

Neto, A. H., & Fiorelli, F. A. S. (2008). Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, *40*, 2169–2176. doi:10.1016/j.enbuild.2008.06.013.

Noh, H. Y., & Rajagopal, R. (2013). Data-driven forecasting algorithms for building energy consumption. In *Sensors and Smart Structures Technologies*

for Civil, Mechanical, and Aerospace Systems (p. 86920T). San Diego: SPIE volume 8692. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84878736246{&}partnerID=tZOtx3y1. doi:10.1117/12.2009894.

Owen, D. B., Abramowitz, M., & Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* volume 7. Courier Corporation. doi:https://doi.org/10.2307/1266136. arXiv:1701.01870.

Papadopoulos, S., Azar, E., Woon, W.-L., & Kontokosta, C. E. (2017). Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *Journal of Building Performance Simulation*, *1493*, 1–11. doi:https://doi.org/10.1080/19401493.2017.1354919.

Park, B., Messer, C. J., & Urbanik II, T. (1998). Short-term freeway traffic volume forecasting using radial basis function neural network. *Transportation Research Record: Journal of the Transportation Research Board*, *1651*, 39–47. doi:https://doi.org/10.3141/1651-06.

Park, Y.-S., & Lek, S. (2016). Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. In *Developments in Environmental Modelling* (pp. 123–140). Wiley Online Library. doi:https://doi.org/10.1016/B978-0-444-63623-2.00007-4.

Paudel, S., Elmtiri, M., Kling, W. L., Corre, O. L., & Lacarrière, B. (2014). Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network. *Energy and Buildings*, *70*, 81–93. URL: http://dx.doi.org/10.1016/j.enbuild.2013.11.051. doi:10.1016/j.enbuild.2013.11.051.

Platon, R., Dehkordi, V. R., & Martel, J. (2015). Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. *Energy and Buildings*, *92*, 10–18. URL: http://dx.doi.org/10.1016/j.enbuild.2015.01.047. doi:10.1016/j.enbuild.2015.01.047.

Rastogi, P. (2016). *On the sensitivity of buildings to climate: the interaction of weather and building envelopes in determining future building energy consumption*. Ph.D. thesis Ecole Polytechnique Fédérale de Lausanne. URL: https://infoscience.epfl.ch/record/220971/files/EPFL{_}TH6881.pdf.

Rastogi, P., & Andersen, M. (2016). Incorporating Climate Change Predictions in the Analysis of Weather-Based Uncertainty. *Proceedings of SimBuild*, *6*.

Rastogi, P., Polytechnique, E., & Lausanne, F. D. (2017). Gaussian-Process-Based Emulators for Building Performance Simulation. In *Building Simulation 2017: The 15th International Conference of IBPSA*. San Francisco: IBPSA.

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk analysis*, *22*, 579–590.

Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering*, *6*, 5. doi:https://doi.org/10.1186/s40327-018-0064-7.

Si, J. (2017). *Green retrofit of existing non-domestic buildings as a multi criteria decision making process by*. Ph.D. thesis.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, *55*, 271–280.

Tin Kam Ho (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (pp. 278–282). IEEE volume 1. doi:https://doi.org/10.1109/ICDAR.1995.598994.

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, *49*, 560–567. doi:10.1016/j.enbuild.2012.03.003.

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption : A comparison of regression analysis , decision tree and neural networks. *Energy*, *32*, 1761–1768. doi:https://doi.org/10.1016/j.energy.2006.11.010.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., & Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. *Energy and Buildings*, *171*, 11–25. doi:https://doi.org/10.1016/j.enbuild.2018.04.008.

Wong, S. L., Wan, K. K. W., & Lam, T. N. T. (2010). Artificial neural networks for energy analysis of office buildings with daylighting. *Applied Energy*, *87*, 551–557. URL: http://linkinghub.elsevier.com/retrieve/pii/S0306261909002669. doi:10.1016/j.apenergy.2009.06.028.

Xifara, A., & Tsanas, A. (2012). Energy efficiency Data Set.

Xuemei, L. X. L., Jin-hu, L. J.-h. L., Lixing, D. L. D., Gang, X. G. X., & Jibin, L. J. L. (2009). Building Cooling Load Forecasting Model Based on LS-SVM. *Asia-Pacific Conference on Information Processing*, *1*, 55–58. doi:10.1109/APCIP.2009.22.

Yalcintas, M. (2006). An energy benchmarking model based on artificial neural network method with a case example for tropical climates. *International Journal of Energy Research*, *30*, 1158–1174. doi:10.1002/er.1212.

Yalcintas, M., & Ozturk, U. A. (2007). An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *International Journal of Energy Research*, *31*, 412–421. doi:10.1002/er.1232. arXiv:arXiv:1011.1669v3.

Yokoyama, R., Wakui, T., & Satake, R. (2009). Prediction of energy demands using neural network with model identification by global optimization. *Energy Conversion and Management*, *50*, 319–327. URL: http://dx.doi.org/10.1016/j.enconman.2008.09.017. doi:10.1016/j.enconman.2008.09.017.

Zhang, Y., O'Neill, Z., Dong, B., & Augenbroe, G. (2015). Comparisons of inverse modeling approaches for predicting building energy performance. *Building and Environment*, *86*, 177–190. URL: http://dx.doi.org/10.1016/j.buildenv.2014.12.023. doi:10.1016/j.buildenv.2014.12.023.

Zhang, Y., O'Neill, Z., Wagner, T., & Augenbroe, G. (2013). An inverse model with uncertainty quantification to estimate the energy performance of an office building. *IBPSA Building Simulation*, (pp. 614–621). URL: http://www.ibpsa.org/proceedings/BS2013/p{_}1410.pdf.

Zhao, H.-x., & Magoulès, F. (2010). Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy Consumption. *Journal of Algorithms & Computational Technology*, *4*, 231–249. doi:10.1260/1748-3018.4.2.231.

Zhao, H. X., & Magoulès, F. (2012a). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*, 3586–3592. doi:10.1016/j.rser.2012.02.049.

Zhao, H.-X., & Magoulès, F. (2012b). Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method. *Journal of Algorithms & Computational Technology*, *6*, 59–77. URL: http://journals.sagepub.com/doi/10.1260/1748-3018.6.1.59. doi:10.1260/1748-3018.6.1.59.