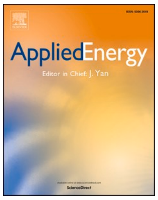
Applied Energy 357 (2024) 122472

Contents lists available at ScienceDirect

Applied Energy

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



Predicting whole-life carbon emissions for buildings using different machine learning algorithms: A case study on typical residential properties in Cornwall, UK

Lin Zheng [a,b], Markus Mueller [b,c,d], Chunbo Luo [d,e], Xiaoyu Yan [a,b,*]

[a] Renewable Energy Group, Engineering Department, Faculty of Environment, Science and Economy, University of Exeter, Penryn Campus, Penryn TR10 9FE, UK
[b] Environment and Sustainability Institute, Faculty of Environment, Science and Economy, University of Exeter, Penryn Campus, Penryn TR10 9FE, UK
[c] Department of Earth and Environmental Sciences, Faculty of Environment, Science and Economy, University of Exeter, Penryn Campus, Penryn TR10 9FE, UK
[d] Institute for Data Science and Artificial Intelligence, University of Exeter, Exeter EX4 4RN, UK
[e] Department of Computer Science, Faculty of Environment, Science and Economy, University of Exeter, Streatham Campus, Exeter, EX4 4RN, UK

HIGHLIGHTS

- Novel method using machine learning algorithms for building WLCE prediction.
- Promotes WLCE integration into the early stages of building design.
- Leverage a comprehensive UK residential dataset for robust model development.

ARTICLE INFO

Keywords:
Whole life carbon emissions
Life cycle thinking
Machine learning algorithms
Sustainable building
Carbon reduction
Data-driven approaches

ABSTRACT

Whole-life carbon emissions (WLCE) studies are critical in assessing the environmental impact of buildings and promoting sustainable design practices. However, existing methods for estimating WLCE are time-consuming and data-intensive, limiting their usefulness in the early building design stages. In response to this, this research introduces a novel approach by harnessing various machine learning algorithms to predict WLCE and WLCE intensity (normalised by floor area) for buildings. To evaluate the suitability of machine learning algorithms, we conducted an experiment involving ten algorithms to build the prediction models. These models were trained using data from 150 typical residential properties in Cornwall, UK, along with 28 features obtained from a comprehensive survey, including floor area, heating type, and occupant characteristics. The ten algorithms include Multiple Linear Regression, and non-linear algorithms such as Decision Tree, Random Forest. Performance evaluation metrics, such as coefficient of determination (R^2), mean absolute error (MAE), means squared error (MSE), root-mean-square error (RMSE), and elapsed time, were employed. Our research contributes to the field by showcasing the effectiveness of machine learning models in predicting building WLCE. We reveal that all the tested machine learning algorithms have the capability to predict WLCE and WLCE intensity, non-linear models outperform linear ones, and the Random Forest (RF) model demonstrates superior performance in terms of accuracy, stability, and efficiency. This research encourages the integration of life cycle studies into the early design stage, even within tight building design schedules, offering practical guidance to architects and designers. Furthermore, these results also benefit a wide range of stakeholders, not only the architects but also the engineers, policymakers, and life cycle assessment (LCA) researchers, contributing to the advancement of data-driven sustainability approaches within the building sector.

Abbreviations: LCA, Life Cycle Assessment; WLCE, Whole-Life Carbon Emissions; $\boldsymbol{R^2}$, Coefficient of Determination; MAE, Mean Absolute Error; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; MLR, Multiple linear regression; LASSO, Least Absolute Shrinkage and Selection Operator; GBR, Gradient Boosting Regressor; DT, Decision Tree; RF, Random Forest; SVR, Support Vector Regression; GPR, Gaussian Process Regression; MLP, Multi-Layer Perceptron; GRNN, General Regression Neural Network; GBDT, Gradient Boosting Decision Trees; CNN, Convolutional neural network.

[*] Corresponding author at: University of Exeter, Penryn Campus, Penryn TR10 9FE, UK.
E-mail address: Xiaoyu.Yan@exeter.ac.uk (X. Yan).

https://doi.org/10.1016/j.apenergy.2023.122472
Received 15 September 2023; Received in revised form 10 November 2023; Accepted 4 December 2023
Available online 20 December 2023

1. Introduction

Buildings are major contributors to global carbon emissions, accounting for nearly 40% of the world's total emissions [1–3]. With buildings exerting such significant pressure on the natural environment, it has become imperative to focus on innovative and sustainable practices that can mitigate their environmental impact. One of the most crucial aspects of sustainable building design and management is WLCE. In recent years, there has been increasing recognition of the importance of predicting WLCE as a vital tool for assessing a building's carbon emissions. WLCE of buildings, an important aspect of life cycle thinking [4], encompasses a holistic approach that goes beyond measuring emissions during a building's operation and extends to the entire lifespan, from the initial construction to its eventual demolition [5]. On the other hand, when designing a building, the initial design stage holds the key to its long-term environmental performance [6]. Effectively quantifying the impact of WLCE is also indispensable in guiding the creation of environment-friendly and sustainable buildings [7–13]. It represents an important instrument in the mission to advance sustainability, inform architectural and engineering practices, and contribute to the global effort to mitigate climate change. This research is a dedicated effort towards these overarching objectives.

The estimation of WLCE for buildings is supported by numerous free or commercial software packages and plug-ins. Some notable examples for the building sector include OneClick LCA [14], IMPACT [15], the Embodied Carbon in Construction Calculator (EC3) [16], and TallyCAT beta [17]. Studies [18–24] focusing on WLCE for buildings have used Excel to construct inventory databases or software tools developed using physics-based simulations for calculations. For example, a WLCE estimator tool for Sri Lankan buildings based on life cycle inventory analysis was created using the Visual Studio C# programming language [25]. In South Korea, efforts were made to facilitate low-carbon building design through the development of a simplified life cycle carbon emission assessment tool [26] and a calculation programme [27]. A computer-based mathematical tool was specifically developed to quantify embodied carbon emissions from steel multi-story structures that comply with British and European standards [28].

These methods have limitations in predicting building WLCE at an early design stage [29]. The early design stage often involves rapid paces of change. Using such methods requires extensive data collection for a life cycle database, which is difficult and time-consuming when designs are changed frequently [30]. Therefore, enhancing the efficiency of life-cycle calculations by reducing the time and data requirements holds the potential to enable designers to assess the environmental impacts of a building at an early design stage. There is an urgent need to develop methods that can predict building WLCE using a limited set of indicators within a short timeframe, ideally a day or two, as opposed to the months-long processes currently in place.

Data-driven approaches, particularly machine learning algorithms, have gained attention in both academic and industrial domains within the sustainability field [31,32]. Life cycle researchers have also identified the potential of data-driven approaches, such as machine learning algorithms, to mitigate the limitations of existing software packages and tools. These methods have also become increasingly popular for establishing prediction models and comparing simulated results in the building sector. However, current building studies [33–39] primarily focus on building energy consumption prediction, while other studies [40–48] focus on building performance analysis (e.g., in terms of thermal comfort, energy efficiency improvement, occupancy evaluation, building material quantities, diagnostics, or control for building HVAC systems) rather than carbon emissions, particularly not from the life-cycle perspective.

However, the researches focus on building energy consumption prediction or performance analysis alone cannot adequately reflect building WLCE due to the lack of consideration for embodied emissions from materials and operational emissions that are influenced by the future energy transition in buildings' whole life cycles. Therefore, there is a lack of comprehensive study on the application of various machine learning algorithms in the field of WLCE for buildings. Furthermore, both WLCE and WLCE intensity (normlised by floor area) need to be considered. Because obtaining WLCE by simply times predicted WLCE intensity with floor area or obtaining intensity by dividing predicted WLCE by floor area, are both not accurate enough to capture the actual value. Therefore, there is also a need to study both WLCE prediction and intensity prediction as well as the comparison between them.

Recently, several studies have employed machine learning algorithms to predict carbon emissions for buildings. For example, F. Pomponi et al. used machine learning models, including RF and ANN, to predict the quantity of building construction materials and integrated the algorithms into the building design software [49]. The results show that the machine learning model demonstrated good performance in estimating the quantity of building materials and can be integrated into the building design software for calculating the embodied emissions from building materials. Mao et al. employed 12 design factors and four machine learning algorithms, including linear regression, MLP, RF, and SVR approaches, to predict average annual WLCE intensity for 207 residential buildings in Tianjin, China [50]. The results demonstrated significant prediction effectiveness, with R^2 values of 0.75 and 0.80 for MLP and SVR, respectively. Ye et al. utilised factors related to occupants and structural characteristics, along with machine learning algorithms like MLR and GRNN, to model energy-related carbon emissions for 294 office buildings in China [51]. The findings indicated that the GRNN model outperformed the MLR model, with a notable impact of the building's structural attributes on energy-related carbon emissions. Ploszaj-Mazurek et al. utilised various machine learning algorithms, including GBR and CNN, to develop supplementary tools for buildings' carbon emissions estimation using simulated data and buildings generated by Grasshopper software that have different shapes and designs [52]. The result showed the tools developed based on machine learning algorithms achieved high accuracy in predicting embodied and operational emissions respectively. Tsay Y-S et al. used the GBDT model to build the life-cycle prediction model for building envelope renovation, taking different renovation strategies simulated by software for a typical residential house [53]. The results show that the GBDT model could predict life-cycle carbon emissions for residential properties.

Overall, the above studies have primarily demonstrated the potential of machine learning algorithms in predicting carbon emissions for buildings though they are subjective to many limitations including limited scope (e.g., some studies have exclusively focused on either operational emissions or embodied emissions in their calculations, neglecting the holistic life cycle perspectives); data source limitations (some studies rely on simulated energy use data generated from software rather than real-world data, which may limit the generalizability of their findings); future carbon intensity overlooked (the potential variations in electricity carbon intensity resulting from future decarbonisation of electricity are often not taken into account); and occupants characteristics neglected (some studies overlook the inclusion of factors related to occupant characteristics which can influence the carbon emissions of buildings). More importantly, there is a notable lack of comprehensive studies dedicated to both the application and evaluation of various machine learning algorithms for predicting building WLCE and WLCE intensity.

To fill the aforementioned gaps and overcome the limitations in existing studies, this paper presents an exploratory study that uses ten different machine learning algorithms to predict the WLCE and WLCE intensity of 150 typical residential properties in Cornwall, UK. Our hypothesis is that machine learning algorithms can be trained using detailed WLCE calculations of typical residential properties in the UK as well as high-level correlated features obtained from a survey of these buildings and their occupants in real-life scenarios. By testing this hypothesis, our objective is to demonstrate the feasibility of using machine learning models to enable rapid and accurate WLCE prediction for individual buildings. If successful, the machine learning model could be able to provide approximate WLCE results for the new building's design instantly based on a description of the new building's characteristics.

This paper represents a significant contribution to the field of sustainable building design and environmental impact assessment. By introducing a novel approach that harnesses machine learning algorithms to predict WLCE and WLCE intensity for buildings, it addresses a critical need for more efficient and data-driven methods in the early stages of building design. The comprehensive evaluation and validation of ten machine learning algorithms, highlights the potential for accurate and efficient WLCE prediction. Beyond its technical contributions, this research encourages the integration of life cycle studies into the early design process, providing practical guidance to architects and designers, even within tight schedules. Moreover, the broader impact of these findings extends to engineers, policymakers, and LCA researchers, fostering the advancement of data-driven sustainability practices within the building sector and ultimately promoting more environmentally responsible construction and design practices.

2. Materials and method

2.1. Overall methodology

The modular solution workflow is presented in Fig. 1. This is a step-by-step workflow [54] for building a machine learning model. The 150 residential properties, including 85 flats, 48 houses, and 17 bungalows, were used as a case study. These properties are typical UK residential properties [55] that were part of the Smartline project, a six-year interdisciplinary research programme [56–58]. We selected these properties because they had high-resolution actual electricity consumption data measured by sensors as well as detailed information on building attributes and occupant characteristics obtained through questionnaires from a comprehensive survey.

The WLCE results, covering the whole building life-cycle from the product stage to the end-of-life stage, were calculated using the structured method described in a previous study [59] and following the guidelines outlined in "Whole-life Carbon Assessment for the Built Environment Guideline from The Royal Institution of Chartered Surveyors (RICS)" [60]. The WLCE intensity results were derived by dividing the calculated WLCE results by the floor area. The methodology for WLCE calculation considered dynamic parameters such as the change in carbon intensity of the national grid under different decarbonisation scenarios in the UK as well as the replacement of building components. The details of the calculation process and assumptions were available in the supplementary materials (see Section S1, Section S2).

The WLCE dataset for model training was therefore completed. The dataset contained the WLCE and WLCE intensity results of 150 case study properties, along with 28 building features. The calculation results were used as the predicted variables for building the prediction models, respectively. The 28 building features from a comprehensive survey are used as predictors. Part of the raw data used in the calculation and modelling is available online at the UK Data Service [61].

After the data pre-processing and feature selection processes, ten different machine learning algorithms were used to establish the WLCE and the WLCE intensity prediction models, respectively. Each prediction model's establishment was independent and did not affect the others. Therefore, for each model training process, the data splitting process that divided the dataset into a training set and a testing set was also independent. The training and testing data during each model establishment process were therefore distinct. For both WLCE and WLCE intensity predictions, the comparison between the actual values and the predicted values of their respective testing sets was carried out for each of the ten prediction models. Residuals for the models, representing the differences between the predicted and true values, were also presented.

To enhance the reliability of the initial results, an additional analysis was conducted. 30 randomly selected case buildings with their features were used as a separate sample. Then the ten prediction models that have been built were tested by the same separate sample, not only using their respective testing sets during the model establishment process. Then the performance evaluation, which included the five metrics for each of the ten models, was carried out. The results of each model represent the best-performing trained model for a specific machine learning algorithm, considering the chosen hyperparameters. The comparison between the results of the WLCE prediction models and the WLCE intensity prediction models was discussed.

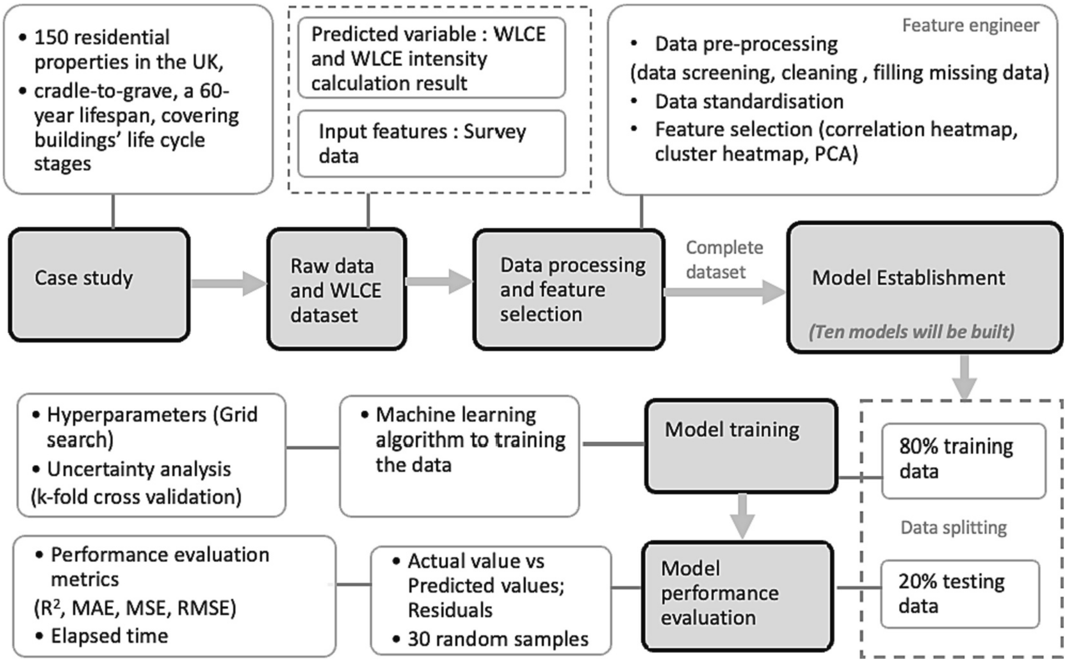

• 150 residential properties in the UK,
• cradle-to-grave, a 60-year lifespan, covering buildings' life cycle stages
Predicted variable : WLCE and WLCE intensity calculation result
Input features : Survey data
Feature engineer
• Data pre-processing
(data screening, cleaning , filling missing data)
• Data standardisation
• Feature selection (correlation heatmap, cluster heatmap, PCA)
Case study
Raw data and WLCE dataset
Data processing and feature selection
Complete dataset
Model Establishment
(Ten models will be built)
• Hyperparameters (Grid search)
• Uncertainty analysis
(k-fold cross validation)
• Machine learning algorithm to training the data
Model training
80% training data
Data splitting
20% testing data
• Performance evaluation metrics
(R², MAE, MSE, RMSE)
• Elapsed time
• Actual value vs Predicted values; Residuals
• 30 random samples
Model performance evaluation

Fig. 1. The solution workflow of prediction model in this paper.

2.2. Data processing and feature selection

Data pre-processing is a technique for converting the data in the initial dataset into a clean and tidy dataset in an appropriate format to conduct further analysis [62]. Organising raw data and conducting appropriate pre-processing are crucial factors in obtaining a reliable data matrix suitable for practical statistical analysis and machine learning [63,64]. The raw data for input features were obtained from a survey that consisted of nearly one hundred questions for occupants of hundreds of residential properties.

To ensure data quality and informative value, we conducted a multi-step data pre-processing approach. This included data extraction and organisation, data cleaning, screening, and selection, dealing with missing data and outliers, and data standardisation. To achieve data standardisation (z-Score normalization) [65] for addressing the issue of imbalanced feature weights, the Standard Scaler method [66] was employed for WLCE dataset. After we completed the standardisation for WLCE dataset, all data in dataset including both calculation results and features were around zero and had a standard deviation of 1. This standardisation made the WLCE dataset suitable for algorithms that rely on the scale of features. The details of data pre-processing approaches, the summary of survey answers, the percentage of missing data for each feature, the summary of continuous data such as energy use and floor area, and the process of standardisation method were provided in the supplementary materials (see Section S2, Section S3, Section S4 and Table S1, Table S2).

The exploratory analysis for feature selection involved various techniques. PCA was employed initially to reduce dimensionality [67,68]. However, due to the nonlinear nature of categorical variables, PCA was used alongside other methods. Spearman correlation analysis was performed and visualised through a correlation heatmap, indicating the strengths and directions of correlations [69]. A clustering heatmap was generated to unveil patterns among features and their relationships with WLCE results, guiding the selection of pertinent features for model creation based on distinct cluster associations [70]. These methods collectively facilitated the identification and curation of relevant features for effective machine learning model construction.

Data processing, standardising, visualisation, and programming such as model training, tuning hyperparameters, and performance metrics evaluation were all implemented in the DataSpell [71] IDE by the Python programme (Python, version 3.9 [72]), using the Scikit-learn [73] and Pandas [74] packages and the Seaborn library [75].

2.3. Machine learning models

For establishing the prediction models, the WLCE dataset was divided into two partitions: 80% for training and the other 20% for testing, as is also often used in building machine learning models [37,43]. Ten parametric and non-parametric machine learning algorithms were used in this paper, including MLR, LASSO Regression, Ridge Regression, GBR, DT, RF, SVR, GPR, MLP, and GRNN. The ten models were built individually for both WLCE and WLCE intensity prediction. MLR was chosen due to its advantages, such as reliability and simplicity of fitting, no hyperparameters required, and no risk of overfitting [40]. Other algorithms are chosen because they have been demonstrated to be suitable for solving regression prediction problems. E.g., DTs have been applied in sustainability studies due to their ability to solve nonlinear relationships and complex interactions [33]. To avoid overfitting, we did not select more complex algorithms for establishing the prediction models due to the limited dataset. The method of building the model using each algorithm is introduced in the subsections below. The supplements to each algorithm, including the objective function and equations, were in supplementary materials (see Section S7).

2.3.1. MLR

MLR, also known as multivariable regression, is a statistical technique that uses several input explanatory variables to predict the outcome of a response variable. Linear regression was the first type of regression analysis to be studied rigorously and is widely used for general-purpose predictive models [76]. The goal of linear regression is to model the linear relationship between the explanatory (independent) variables and the response (dependent) variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression [77]. The MLR method is simple to apply, easy to interpret, and, in some cases, can outperform non-linear models when the sample is small [40]. In linear regression, the relationship is modelled with a linear predictor function whose unknown model parameters are estimated from the data, and such models are called linear models [78].

2.3.2. LASSO regression

LASSO regression models, also known as regularisation methods, are regression analysis methods that are widely used in machine learning and were well suited for models with high degrees of multicollinearity or to automate specific steps in model selection, such as feature selection and parameter elimination [79]. Compared to linear regression, Lasso regression could enhance the prediction accuracy and interpretability of the resulting statistical model. Lasso regression performs L1 regularisation, which adds a penalty equal to the absolute value of the magnitude of the coefficients. Simple, sparse models with few coefficients could be well generated by the Lasso approach (e.g., models with fewer parameters). Some coefficients could become zero and be removed from the model. Higher penalties provide coefficient values that are closer to zero, which is advantageous when creating more straightforward models [80–82].

2.3.3. Ridge regression

Ridge regression is a straightforward predictive algorithm that can be easily formulated and solved [83]. In Ridge regression, a new function is generated by combining the sum of the squared estimate of the error function and the penalty value with the number of parameters, which is used to estimate the parameters of the regression model.

2.3.4. DT

DT is a simple and understandable non-parametric machine learning tool that is used for both regression and classification questions [84]. A decision tree builds regression or classification models in the form of a tree structure. It incrementally develops an associated decision tree while segmenting a dataset into smaller and smaller sections [85]. Decision trees usually consist of three nodes (leaf nodes, internal nodes, or root nodes) and branches (bifurcations). The original node, known as the root node, which represents the complete sample, may be further subdivided into further nodes. Branching represents the decision criteria, whereas internal nodes indicate the dataset's features. The outcomes are finally represented by the leaf nodes. The complexity of the model is influenced by the branches.

Due to their advantages, including effectiveness, requiring less data, and interpretability, DT algorithms have been applied in multiple fields [86]. While overfitting issues can occur with large datasets, we chose to build the simpler DT model in this paper because our dataset is relatively small and avoids the overfitting problem.

2.3.5. RF

RF, a powerful ensemble machine learning algorithm, combines multiple DTs to efficiently classify or predict outcomes. It is popular for its flexibility, practicality, and effectiveness in handling high-dimensional data [50,87]. RF builds numerous decision trees in parallel, effectively reducing both the bias and variance of the model [88]. After generating a large number of trees, they collectively vote for the most popular class. These procedures are commonly referred to as RFs RF uses random sampling of training observations and random subsets of candidate variables for splitting nodes, which is one of the fundamental differences between RF and DT.

2.3.6. GBR

The GBR and RF algorithms are comparable. However, in contrast to RF, the GBR algorithm constructs each tree based on the results of the preceding trees, aiming to identify weights larger than those discovered in the preceding trees [33,83]. Typically, GB generates a set of weak or moderate predictors and leverages their strengths [89]. This algorithm has recently been employed in several studies on the building energy and carbon emission sectors, such as [33,52,83] and has exhibited acceptable prediction accuracy.

2.3.7. SVR

Super vector machine is a learning algorithm that has gained significant popularity in solving classification problems. The sparse solution and good generalisation of the super vector machine lend themselves to adaptation to regression problems [90]. The SVR technique can be employed to investigate the regression relationship between independent variables and continuous dependent variables, making it a valuable tool for solving forecasting problems [91]. The objective function and constraints of SVR aim to find the optimal values of the weight vector w and bias term b that minimise the loss while satisfying the specified tolerance and non-negativity conditions.

2.3.8. GPR

The GPR model is based on Gaussian processes, which enable flexible probabilistic predictions and uncertainty estimation [92] for regression. Gaussian processes are stochastic processes where any finite set of random variables follows a joint Gaussian distribution [93]. The GRP model represents the relationship between input variables and output variables as a function distribution, allowing for uncertainty estimation in predictions. The GPR model is particularly useful in handling small to medium-sized datasets or when the data contains noise or uncertainty. They can handle nonlinear relationships and provide not only point estimates but also uncertainty measures for each prediction [94].

2.3.9. MLP

The MLP algorithm is a type of feed-forward artificial neural network. It is a mathematical model that simulates the actions of brain neurons using fixed layers and neurons [51,95]. A standard artificial neural network architecture consists of input, output, and hidden layers. The input layer takes all the input values, while the output layer generates the result. The presence of hidden layer(s) essentially guarantees that artificial neural network models have non-linear relationships due to intervention between the input and output neurons [33].

2.3.10. GRNN

GRNN is an enhanced neural network technique based on nonparametric regression neural networks, specifically designed for regression problems. It is based on the radial basis function network architecture and is known for its ability to handle non-linear relationships in data. The GRNN essentially functions as a neural network-based function consisting of four layers: input, pattern, summation, and output [51]. The network can forecast the output or outcomes of feeding it a fresh test data set after being trained with the training data set [96]. Back-propagation neural networks often struggle to obtain sufficient data from operating system measurements [97]. Therefore, the use of GRNN is particularly beneficial as it can predict outcomes using only a limited number of training samples. Additionally, GRNN requires a significantly lower amount of additional knowledge to achieve satisfactory fitting [96]. In the GRNN, the output is estimated using the weighted average of the outputs from a training data set. The Euclidean distance between the training data and the test data is used to calculate the weight [97].

2.4. Uncertainty analysis

Cross-validation [98], as the most common approach to ensuring the robustness of the model, was adopted to conduct the uncertainty analysis of models in this paper. In the basic approach, called k-fold cross-validation, the training set is split into k smaller partitions. The following procedure is followed for each of the k “folds”: A model is trained using the folds as training data, and the resulting model is validated on the remaining part of the data. The performance measure reported by k-fold cross-validation is calculated as the average of the values computed during the loop (see Eq. (1)), and commonly, k is set to 5 or 10 [98–100]:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i \tag{1}$$

In this paper, k-cross validation was applied to all the machine learning algorithms. The random split process was repeated five times during cross-validation to analyse the sensitivity of the results.

2.5. Hyperparameter Tunning

A grid search with 5-fold cross-validation is employed to select the optimal parameters for each algorithm, except the MLR. This approach involves dividing the training data into five equally sized folds, randomly shuffling them, and iteratively training the model on nine folds while evaluating its performance on the remaining fold. By carefully tuning the model using this procedure, the risk of encountering fitting problems is minimised. The best score function is utilised to determine the combination of hyperparameters and associated arguments that yield the highest mean cross-validation score, thereby indicating the optimal model configuration. The mean cross-validation scores serve as a reliable basis for selecting the most suitable hyperparameter combination across all models [84].

2.6. Performance evaluation of each model

2.6.1. Predicted value vs actual value, residuals of each model

In terms of performance evaluation for each model, a comprehensive analysis was undertaken. A key aspect of this analysis involved the generation of scatter plots, meticulously illustrating the linear regression between actual values and predicted values from each of the ten models. These scatter plots are recognised as vital tools in the assessment of model performance [101]. The overall trend of the points on the scatter plot indicates the goodness of fit, with a straight line suggesting a strong alignment between the actual and predicted values. The maximum error distance shows how well the regression lines align with actual values, while the proximity of data points to the ideal path ($y = x$) demonstrates the accuracy of predictions. The noteworthy performance in predicted values compared to actual values would be attributed to the impact of K-fold cross-validation [102].

The residuals plots were used to show the residuals between the actual value and the predicted value. A residuals plot is a diagnostic tool that helps assess how well a machine learning model fits the data and whether there are any patterns or errors in the predictions [103]. In a residuals plot, the x-axis typically represents the predicted values (or fitted values) generated by the machine learning model. The y-axis represents the residuals, which are the differences between the actual observed values and the predicted values for each data point [104]. A random scattering of residuals around zero is indicative of a well-performing model, suggesting that its predictions are unbiased. Conversely, the presence of a discernible pattern in the residuals plot, such as a curve or funnel shape, may signal that the model is failing to capture a non-linear relationship within the data. When the residuals exhibit a systematic trend, such as an increase or decrease with changing predicted values, it points to the existence of bias or systematic errors within the model. The identification of outliers within the residuals plot may signify data points that the model struggles to handle effectively.

2.6.2. Validation testing using a separate dataset

To strengthen the reliability of our initial findings, we performed an additional validation test using a separate dataset that included thirty randomly chosen case buildings. All ten prediction models were collectively evaluated using this dataset. We illustrated the predicted values and actual values for each of the thirty residential properties in ten scatter plots, one for each model. The primary objective of this validation test was to assess the models' ability to generalise, ensuring they performed well not only on their training and testing data (comprising 80% and 20% of the WLCE dataset) but also on the same separate dataset. By using this common dataset for all models, we gained a comprehensive understanding of their performance and adaptability.

2.6.3. Performance evaluation metrics

All the prediction models were assessed to evaluate their performance. The elapsed time for the algorithms to train the data and establish the model was used to compare their efficiency. We employed three widely accepted evaluation metrics to assess the performance of our predictive models [105]: R^2, MAE, MSE, and RMSE. R^2 is a metric used to quantify the goodness of fit of a regression model. It measures the proportion of the variance in the dependent variable (WLCE or WLCE intensity in our case) that can be explained by the independent variables (features used for prediction) and offers insights into the predictive power of our models. MAE, as the name suggests, calculates the mean of the absolute differences between the predicted and observed values, providing a measure of the average prediction error. MSE quantifies the average of the squared differences between predictions and actual values, giving higher weight to large errors. RMSE represents the square root of MSE and is useful for providing an interpretable measure of prediction accuracy in the original units of the data. The equations [37] for the calculation of the R^2, MAE, MSE, and RMSE are as below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\overline{y} - y_i)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}} \tag{5}$$

These metrics were suitable for regression analysis tasks and were able to provide a comprehensive evaluation of our models, capturing aspects of accuracy, precision, and overall goodness of fit. Eq. (2) calculates the R^2. The R^2score can range from 0 to 1, where a score of 1 indicates a perfect fit, meaning that the model's predictions match the observed values exactly. Conversely, a score of 0 suggests that the model provides no better prediction than simply using the mean of the observed values. Eq. (3) calculates the MAE, which is the average of the absolute differences between the predicted $\widehat{y}$ and observed $y$values. This metric provides a measure of the average prediction error, where a lower MAE indicates better predictive accuracy. MAE is particularly useful when the magnitude of errors should be directly interpretable. Eq. (4) represents the MSE, which computes the average of the squared differences between predicted $\widehat{y}$ and actual y values. MSE places more weight on larger errors and is often used to penalise models for larger deviations from actual values. Eq. (5) calculates the RMSE, which is the square root of the MSE. RMSE is helpful because it provides a measure of prediction accuracy in the original units of the data. It is particularly useful when an interpretable measure is needed for assessing how well the model fits the observed data.

3. Results and discussion

3.1. WLCE calculation results

Fig. 2 (a) displays the calculation results of WLCE and WLCE intensity for the 148 residential buildings. Two out of the 150 case properties were removed during data pre-processing due to lower data quality. The left vertical axis represents the WLCE results, which exhibit a range of 20 to 195 t with an average of 84 t for the sampled residential properties. Simultaneously, the right axis represents the WLCE intensity, indicating values within the range of 0.3 to 2 t per square metre, with an average of 1 t per square metre.

Fig. 2 (b) shows the visualisation histogram and bivariate distribution of standardised WLCE and WLCE intensity calculation results. Because of the standardisation for WLCE dataset, all data including both the calculation results and features were around zero and had a standard deviation of 1. As shown in Fig. 2 (b), the y-axis is the standardised WLCE and its intensity calculation result. The histogram reveals the spread of values for both WLCE and WLCE intensity. It is apparent from this distribution that both WLCE and WLCE intensity values closely approximate a normal distribution. The bivariate distribution plot allows us to explore potential correlations or patterns between these two factors. The fact that both distributions approximate a normal shape suggests that our selection of case buildings represents a diverse yet representative sample of residential properties. This reinforces the robustness of our dataset and its suitability for further predictive modelling. These visualisation results set the stage for our in-depth exploration of machine learning algorithms' effectiveness in predicting WLCE and WLCE intensity for the betterment of environmentally conscious building design.

Results related to data organisation (see Table S1 and Table S2), pre-processing, and feature selection, which encompass the correlation heatmap (see Fig. S2) and clustering heatmap (see Fig. S3) depicting the relationship between WLCE and the features, can be found in the supplementary materials.

3.2. Prediction model results

3.2.1. Predicted value vs actual value, residuals of each model

Fig. 3 (a) displays the comparison of actual and predicted values for the testing set, which constitutes 20% of the total dataset and includes 30 data points. It also shows the maximum error for each WLCE prediction model during model-building. The black scatters are the coordinates of the actual and predicted standardised WLCE value (per tonne of carbon emissions), and the red line is the fitting beeline of the black scatters. The x-axis shows the predicted WLCE value (standardised), and the y-axis shows the actual WLCE value (standardised). The maximum error between the actual value and the predicted value was also presented by the red dotted line. As shown in Fig. 3 (a), it can be observed that for all models, the actual values increase with the predicted values. The linearity of the points around the 45-degree line suggests a good model fit for each regression model. These results indicate a strong correlation between the predicted values and the actual values for all models.

To further evaluate the performance of the models, Fig. 3 (b) presents the plot of residuals for each algorithm. Fig. 3 (b) depicts the residuals from each model, which are important for diagnosing model fit and identifying potential data issues. The residuals represent the differences between the predicted and actual WLCE values. As shown in Fig. 3 (b), it is evident that the residual points in each model graph exhibit a random

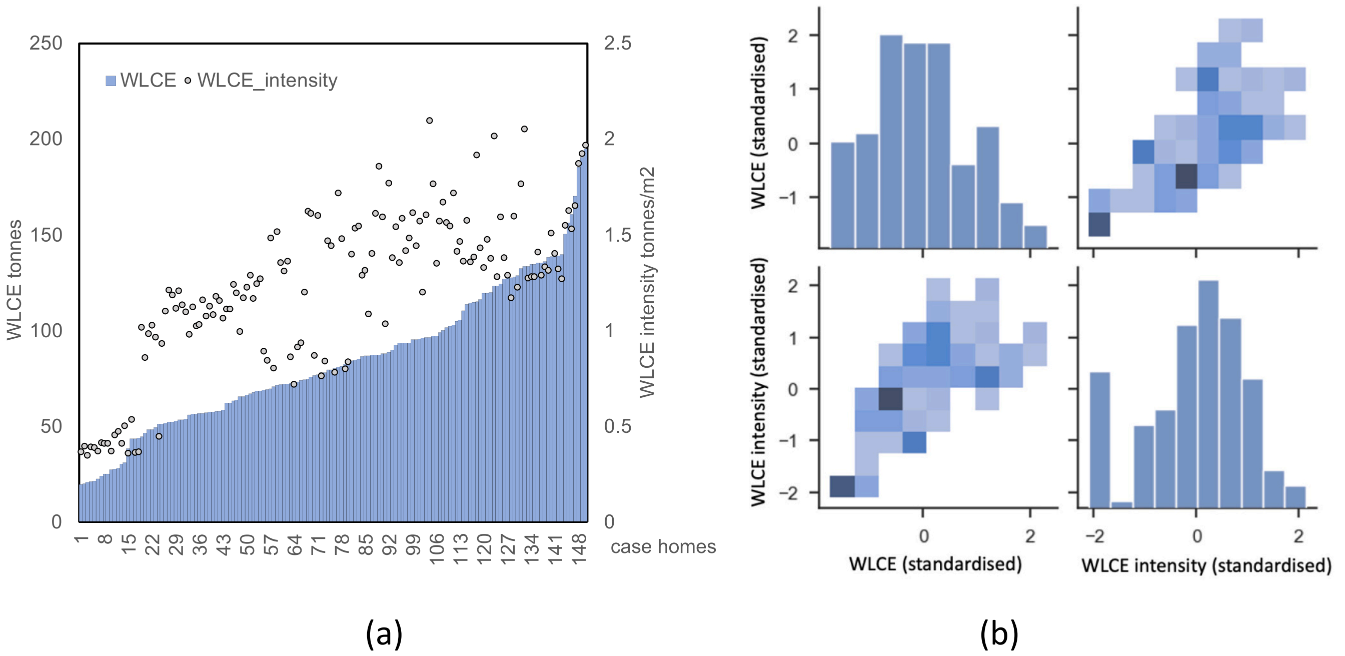


Fig. 2. (a) Results of WLCE and WLCE intensity calculations; (b) Visualizing histogram and bivariate distribution of standardised WLCE and WLCE intensity.

and uniform distribution within the range of −2 to 2. The distribution of the residual points does not display any discernible patterns or rules. Therefore, the residual results also demonstrate the feasibility of machine learning models to approximately predict the WLCE. These results provide evidence to support the initial hypothesis of this study that establishing WLCE prediction models using machine learning algorithms is indeed feasible.

3.2.2. Validation testing using a separate dataset

Fig. 4 displays the test results using a separate dataset containing the same 30 random properties for ten WLCE prediction models. This figure extends the analysis of predictive performance across ten WLCE prediction models. The x-axis represents the 30 case properties; blue dots indicate actual standardised WLCE values, and orange dots represent predicted values by the different model. Lines connect each dot to facilitate better visual comparison. Fig. 4 clearly shows that, for this same group of 30 properties, the actual and predicted values of each property by each model are quite close. This consistency in results, as also seen in Fig. 3, highlights that different machine learning methods were able to provide reasonably accurate predictions for the WLCE results of these specific properties.

The prediction outcomes for WLCE intensity also align with these two aspects: (1) the comparison of actual and predicted values, including residuals for each model, and (2) validation testing using a distinct dataset comprising 30 randomly selected residential properties. To prevent redundancy in presenting similar figures, these results related to intensity prediction are provided in the supplementary material (see Figs. S4 and S5).

3.2.3. Performance evaluation metrics of both WLCE and intensity prediction models

Table 1 provides the overall results of the performance metrics evaluation of the ten models for both WLCE and WLCE intensity performance. For WLCE prediction models, the RF, SVR, and MLP models demonstrate superior predictive performance in forecasting WLCE for residential properties compared to the others, particularly the linear regression models (MLR, LASSO, and Ridge), based on evaluation scores. Among these three algorithms, the RF model outperforms the others in terms of R^2 (0.94 for the train set, 0.89 for the test set) higher than others. MAE (0.16 for the train set, 0.23 for the test set), MSE (0.05 for the train set, 0.09 for the test set), and RMSE (0.23 for the train set, 0.30 for the test set) are lower than others. SVR algorithm ranks second, followed by the MLP algorithm in third place.

For WLCE intensity prediction models, the RF, GBR, and GRNN models demonstrate superior predictive performance compared to the others, particularly the linear regression models (MLR, LASSO, and Ridge). Among these three algorithms, the RF model also outperforms the others in terms of R^2 (0.87 for the train set, 0.85 for the test set) higher than others. MAE (0.24 for the train set, 0.29 for the test set), MSE (0.13 for the train set, 0.12 for the test set), and RMSE (0.37 for the train set, 0.35 for the test set) are lower than others. The GBR algorithm ranks second, followed by the GRNN algorithm in third place.

In terms of elapsed time, although the RF algorithm is not the fastest among the top algorithms for both WLCE and WLCE intensity prediction tasks, it is still considered to be the most favourable choice due to the negligible time difference. However, for larger datasets, more attention might need to be given to elapsed time when selecting or comparing these algorithms. The hyperparameters for each WLCE and WLCE intensity prediction models were available in supplementary materials (see Table S3).

3.3. Discussion and limitations

3.3.1. Discussion

In our study, for both WLCE and WLCE intensity, all ten models have the capability to predict the results accurately and efficiently. Non-linear regression models such as RF, GBR, and SVR have better performances than linear regression models such as MLP, indicating the non-linear relationship between the features and the calculation results. The normality of the WLCE and WLCE intensity calculation results (see Fig. 2) underscores the significance of our selection of case buildings, which captures a representative sample of residential buildings in terms of their WLCE and WLCE intensity.

There are some commonalities and noticeable differences between the WLCE and WLCE intensity prediction results. RF, SVR, and MLP are the top-performing models for WLCE prediction, while RF, GBR, and GRNN are the top three models for WLCE intensity prediction. Therefore, RF performs the best for both WLCE and WLCE intensity

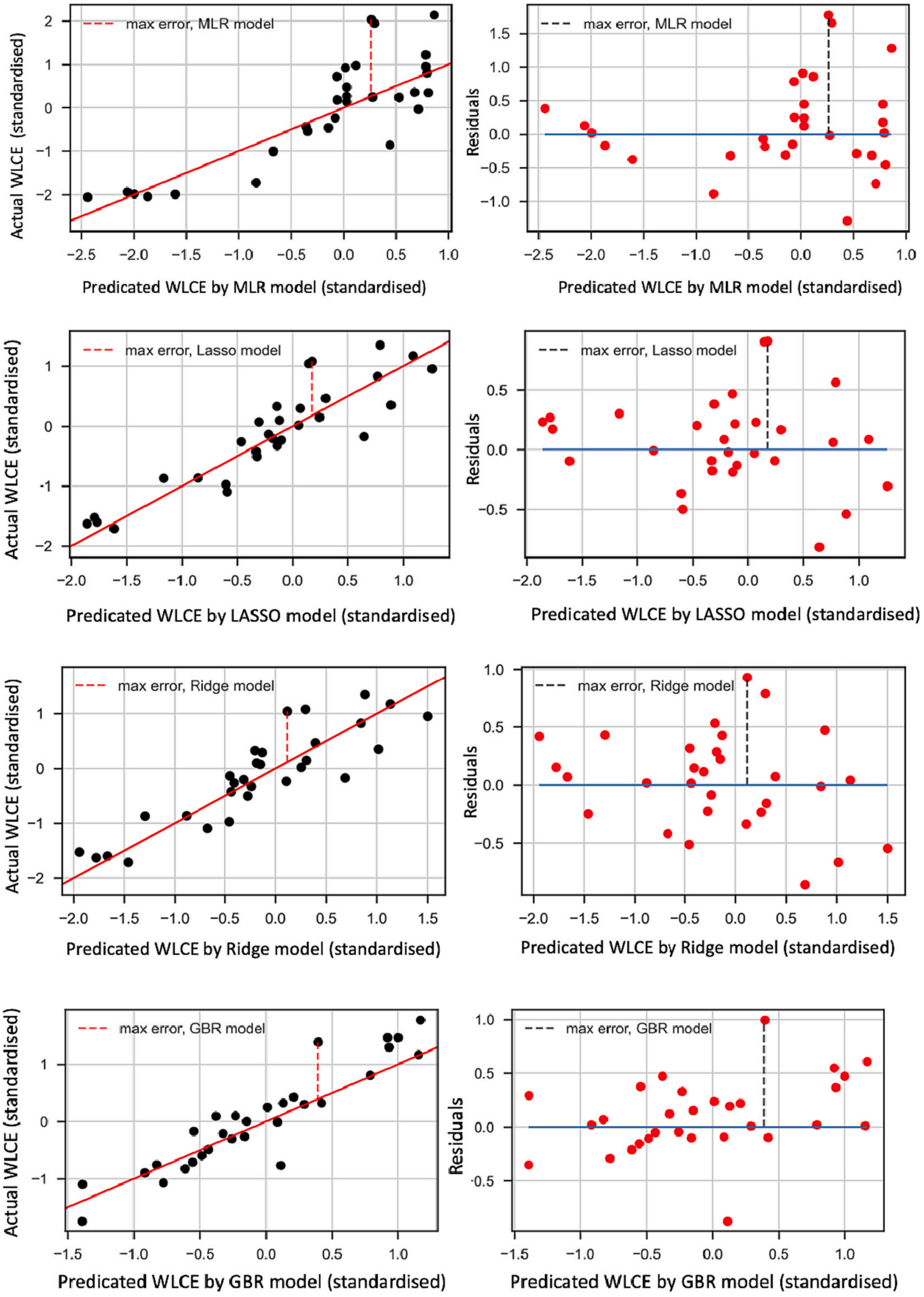


Fig. 3. Comparison of actual versus predicted values and analysis of residuals for standardised WLCE per tonne of carbon emissions using different regression models. (a) Actual vs. Predicted values for standardised WLCE (per tonne carbon emissions); (b) Residuals between true and predicted values for standardised WLCE.

predictions. The RF model outperformed in terms of accuracy, stability, and efficiency for both WLCE and WLCE intensity prediction. The RF algorithm systematically permutes and tracks the decrease in prediction accuracy for each, assigning relative importance scores to individual variables [33,106]. This process ensures the robustness of the RF algorithm by avoiding overfitting and emphasising the overall predictive power rather than fixating on specific predictor-response relationships.

The results from another study [84] also indicate that RF performs better than DT and SVR when predicting carbon emissions associated with the whole building sector in China. However, another study [50] showcased the predictive strength of SVR ($R^2 = 0.80$) in forecasting WLCE intensity for residential properties in Tianjin, China,

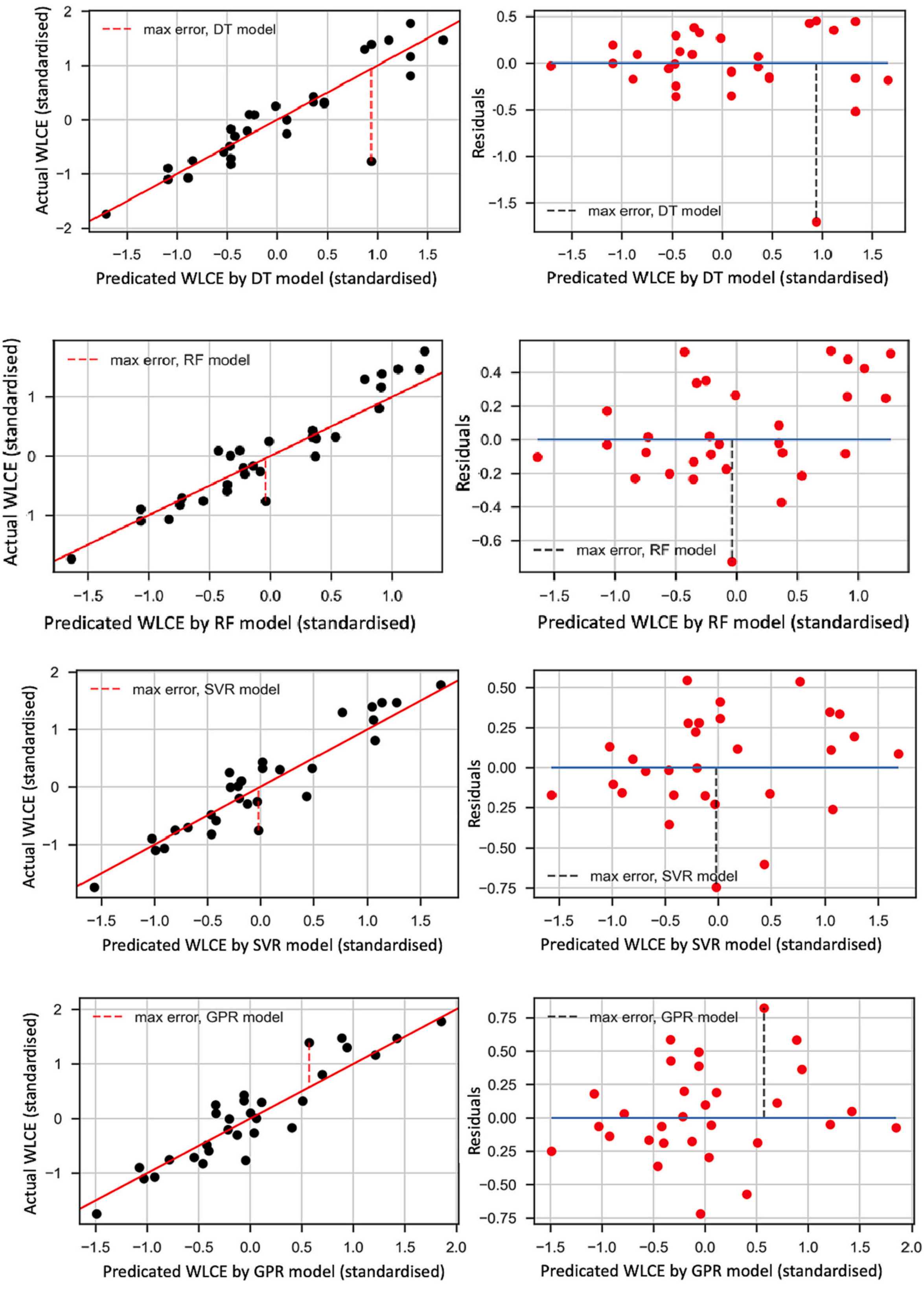
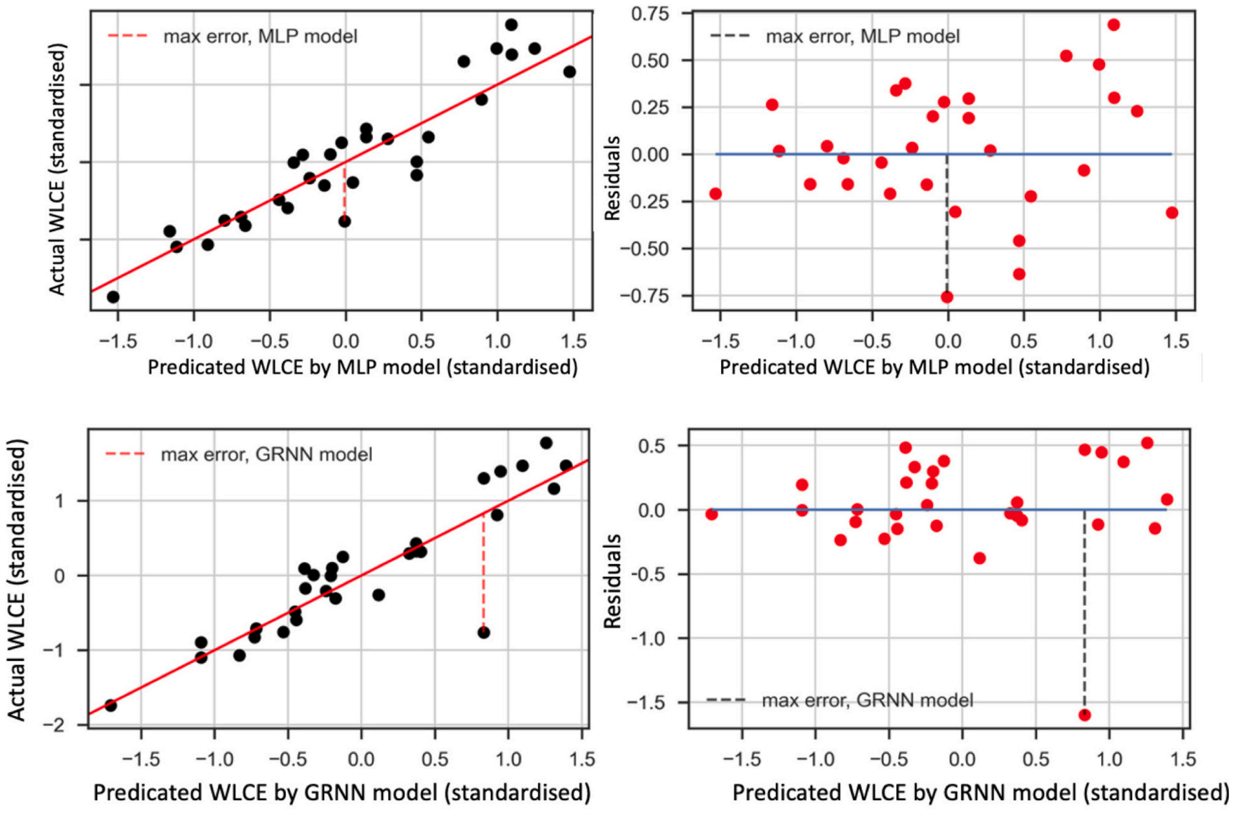


Fig. 3. (*continued*).

outperforming RF ($R^2 = 0.65$). In this paper, RF showed slightly better predictive capabilities than SVR for forecasting WLCE and WLCE intensity for residential properties in Cornwall, UK (see Table 1). Despite both studies focusing on residential properties, the difference in results can be attributed to several factors, including variations in input features, the predicted variable, data quality, and the geographical locations of the properties. Notably, our study employed actual electricity consumption data rather than simulated data generated by software.



(a) (b)

Fig. 3. (*continued*).

It is also worth mentioning that while GRNN ranks as the third-best model for predicting WLCE intensity, it does not feature among the top three models for WLCE prediction. Consistent with our findings, another study [51] demonstrates that GRNN outperforms MLR ($R^2 = 0.76$ vs. $R^2 = 0.70$) when predicting annual carbon emissions intensity for office buildings. Potential applications of machine learning algorithms in life cycle analysis studies are numerous; however, the choice and utility of specific algorithms for the project will depend on the aim, the type of data, and the dataset's size [4]. Our study could provide valuable insights into evaluating the efficacy and suitability of diverse machine-learning algorithms.

We compared the performance evaluation results in Table 1 for the prediction of WLCE and WLCE intensity using the same algorithms. It can be concluded that machine learning algorithms generally perform better in predicting WLCE compared to WLCE intensity. For example, in terms of WLCE prediction, RF achieves higher accuracy with an R^2 of 0.94 for the training dataset and 0.89 for the test dataset. In contrast, for WLCE intensity prediction, the corresponding R^2 values are lower at 0.87 for the training dataset and 0.85 for the test dataset.

The observed difference in WLCE prediction versus WLCE intensity prediction can be attributed to the role of the “floor area” factor. While the “floor area” feature appears to be important in predicting WLCE, it has less correlation with the intensity results and exerts minimal influence on the prediction results for WLCE intensity (see Fig. S1). The reason behind this discrepancy could be that WLCE considers the overall carbon emissions associated with a building, which can be influenced by various factors, including the floor area. However, WLCE intensity focuses on the carbon emissions per unit of area, thereby potentially diluting the impact of the “floor area” factor.

Consequently, when predicting WLCE, the better performance of machine learning models can be attributed to the prominence of the “floor area” factor. Other features might have stronger associations with WLCE intensity, leading to differences in feature selection and predictive performance between the two predicted variables. This result suggests different scenarios for practical applications. For example, in the early design stages of a building, the floor area varies significantly across different architectural designs. When using floor area as a feature for the WLCE prediction, architects or engineers can observe the differences in WLCE among different designs more intuitively and accurately. In addition, in the context of calculating building carbon footprints, predicting WLCE intensity can provide a quicker approximation of the carbon footprint per unit of building area, improving the efficiency of the life cycle estimation.

3.3.2. *Limitations and future work*

The limited sample size restricts the investigation of the applicability of these different models to larger datasets. Future research should therefore expand the sample size of case properties to test the generalizability of the findings of this study. Further research could also explore some additional key variables, particularly regarding the construction structures, materials, and product use, load characteristics of household appliances or electrical equipment, thermal performance, and occupant behaviour, which could potentially help machine learning algorithms generate more reliable predictions of building WLCE. The methods for improving the robustness, stability, and performance of the different prediction models could be studied in future studies.

The calculation of WLCE involves several assumptions that contribute to the inherent uncertainty associated with the utilisation of materials or products. Our case buildings were residential properties located in Cornwall, UK, and therefore had similar climate conditions. Future research could also study whether various machine learning algorithms are useful and effective in predicting the WLCE of buildings of other types and in different climate zones or conditions. In addition, the operational emissions were influenced by the energy mix structure in

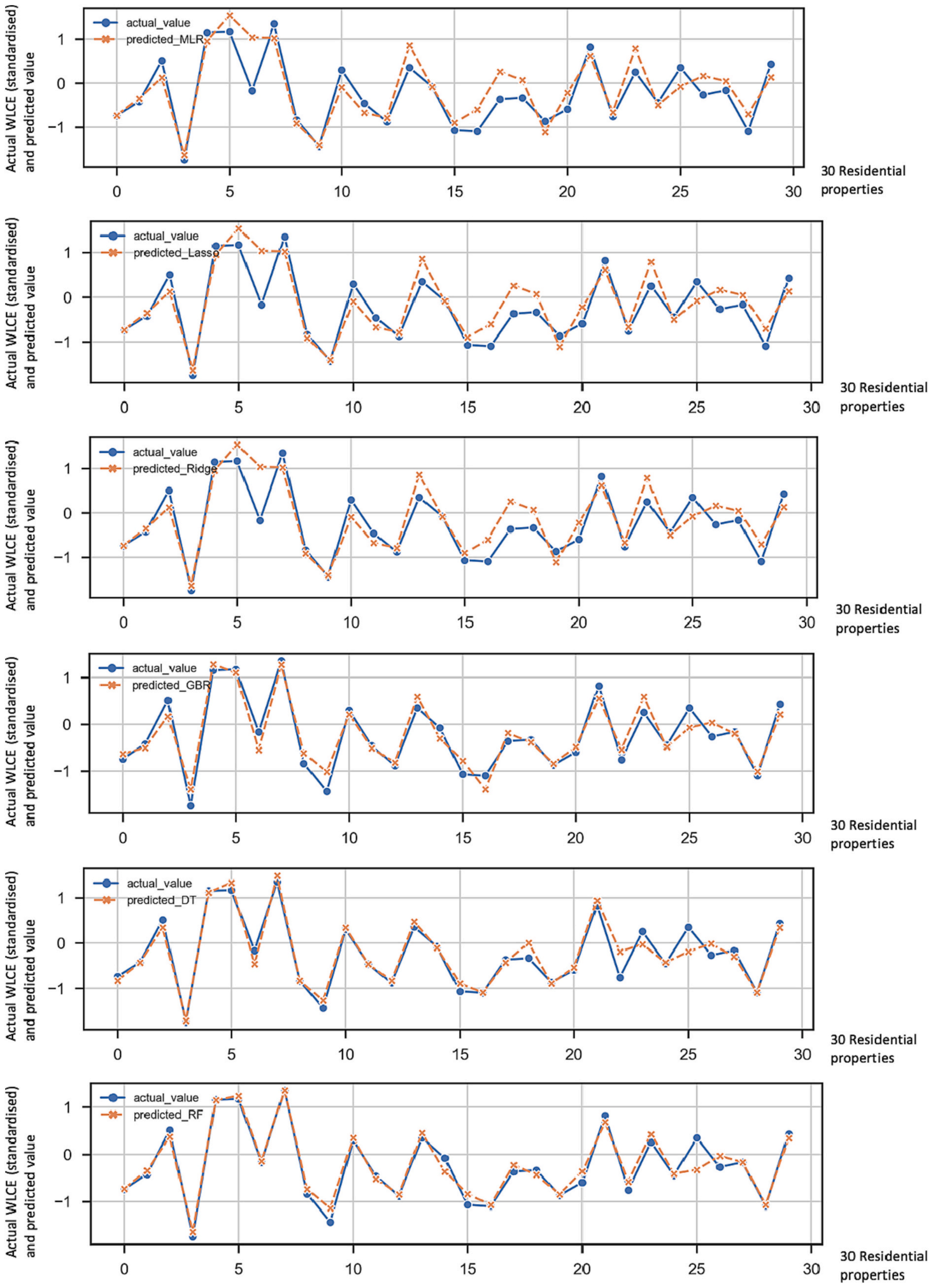


Fig. 4. Scatter plot of each model: actual vs. predicted values for 30 properties. The x-axis represents the 30 case properties; blue dots indicate actual standardised WLCE values, and orange dots represent predicted values by the different model; Lines connect each dot to facilitate better visual comparison. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

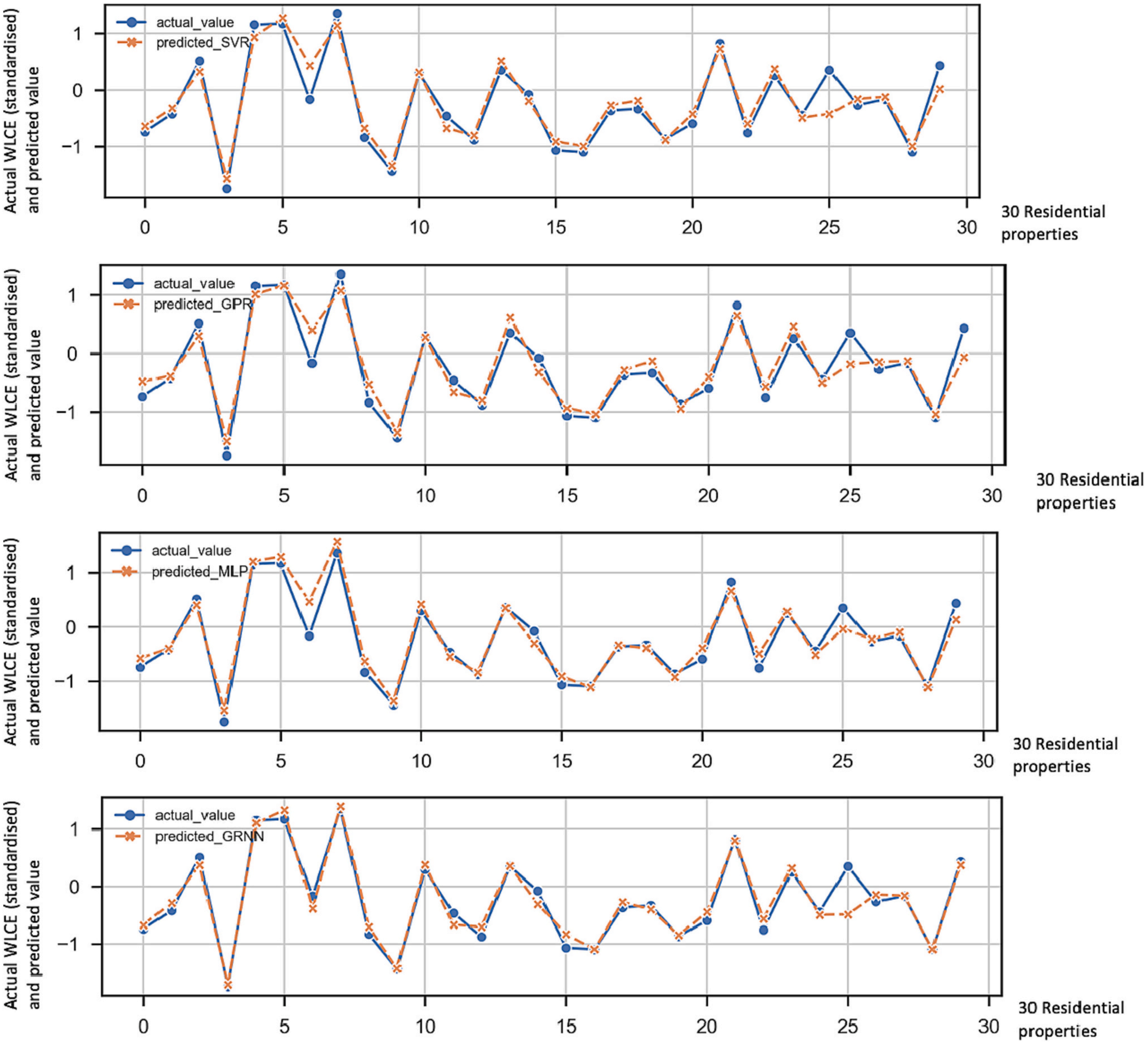


Fig. 4. (*continued*).

different countries and the relevant policies during the buildings' whole life cycle, generally over 50 years.

Despite its limitations, the results indicate the application of machine learning algorithms to predict buildings' WLCE is a promising research area. The proposed prediction models for buildings' WLCE could be expanded beyond current scope by establishing benchmarks for different types of buildings and including additional factors in the models. The factors could include building locations (spatial considerations), temporal aspects related to the energy transition and indoor environmental quality scenarios throughout the building's life cycle, and social considerations like occupants' behaviour and their perception of thermal comfort. This would require interdisciplinary studies involving computer science, statistics, big data technology, HVAC engineering, human behaviour, and psychology. Undertaking such interdisciplinary studies poses challenges, but it offers the potential for significant advancements in understanding and addressing the complexities of sustainability in the building sector by advanced data-driven tools.

4. Conclusion

This study explored the application of ten different machine learning algorithms for predicting building WLCE and WLCE intensity. To conclude, our study makes notable contributions to the fields of building sustainability and life cycle assessment. We establish that machine learning algorithms offer an effective and efficient means of predicting WLCE and intensity. By comparing ten different algorithms, we demonstrate that non-linear models, particularly the RF model, outperform linear models. This finding opens the door to data-driven decision-making in early building design and WLCE research. In essence, our study underscores the potential for machine learning to enhance the environmental performance of building designs and offer valuable insights to researchers and practitioners.

In particular, the machine learning algorithms can predict the WLCE and the WLCE intensity results accurately using features related to the building attributes and occupant characteristics. Compared to existing calculation methods, the machine learning models can process and analyse large amounts of data quickly, thereby increasing the availability of WLCE assessments of a large number of buildings together

Table 1
Performances evaluation for the WLCE prediction and WLCE intensity prediction models.

Modelling algorithms	Prediction models	R^2		MAE		MSE		RMSE		Elapsed time (seconds)
		Train	Test	Train	Test	Train	Test	Train	Test	
MLR	WLCE	0.82	0.75	0.30	0.35	0.16	0.18	0.40	0.43	0.00279
	WLCE Intensity	0.68	0.68	0.37	0.50	0.26	0.48	0.51	0.49	0.00411
LASSO	WLCE	0.79	0.80	0.31	0.29	0.18	0.14	0.43	0.38	0.00145
	WLCE Intensity	0.68	0.69	0.38	0.49	0.27	0.46	0.52	0.68	0.00209
Ridge	WLCE	0.81	0.77	0.30	0.33	0.16	0.17	0.40	0.41	0.00217
	WLCE Intensity	0.68	0.67	0.38	0.51	0.27	0.48	0.52	0.69	0.00405
GBR	WLCE	0.91	0.83	0.19	0.26	0.08	0.13	0.28	0.36	0.0150
	WLCE Intensity	0.84	0.84	0.27	0.30	0.15	0.13	0.39	0.36	0.0191
DT	WLCE	0.93	0.82	0.16	0.26	0.06	0.14	0.25	0.37	0.00245
	WLCE Intensity	0.81	0.71	0.28	0.34	0.19	0.23	0.44	0.48	0.0019
RF	WLCE	0.94	0.89	0.16	0.23	0.05	0.09	0.23	0.30	0.0472
	WLCE Intensity	0.87	0.85	0.24	0.29	0.13	0.12	0.37	0.35	0.1542
SVR	WLCE	0.88	0.88	0.19	0.25	0.10	0.09	0.32	0.30	0.0025
	WLCE Intensity	0.79	0.81	0.27	0.29	0.21	0.15	0.46	0.39	0.00227
GPR	WLCE	0.92	0.85	0.18	0.26	0.07	0.12	0.27	0.34	0.0718
	WLCE Intensity	0.76	0.66	0.35	0.44	0.23	0.28	0.48	0.52	0.0139
MLP	WLCE	0.92	0.86	0.16	0.27	0.07	0.11	0.26	0.33	0.188
	WLCE Intensity	0.73	0.67	0.40	0.44	0.27	0.26	0.52	0.51	0.1625
GRNN	WLCE	0.89	0.81	0.18	0.25	0.09	0.15	0.31	0.38	0.00186
	WLCE Intensity	0.85	0.81	0.22	0.30	0.15	0.40	0.39	0.40	0.00083

and/or at the early design stage of buildings. Moreover, the machine learning algorithms can automatically learn from and improve upon data and be adapted to suit a variety of different datasets, allowing more potential studies on different predictors related to buildings, occupants, and the environment (e.g., climate conditions).

These findings have practical implications for decision-making and resource optimisation in different contexts. By identifying the factors that have the greatest impact on building WLCE, the machine learning prediction models could support data-driven decision-making in applying life cycle thinking at the early design stage despite the tight building design schedule, helping architects and engineers improve the environmental performance of the building designs. The prediction models can also assist researchers in obtaining quick approximations of the carbon footprint of buildings, thereby enhancing the efficiency of time-consuming building WLCE research. Overall, this study suggests that machine learning algorithms can be a potentially viable alternative to current tools or an addition to the toolkits to provide fast yet high-quality forecasts of building WLCE and WLCE intensity for practitioners, researchers, and policymakers.

Author contribution

Lin Zheng contributed to the study design, formulated the research questions, outlined the methodology, data collection, calculation, modelling, analysis, and interpretation of results, and wrote the manuscript. Xiaoyu Yan guided the overall methodology and interpretation of the results and critically reviewed and proofread the manuscript for clarity and accuracy. Markus Muller and Chunbo Luo also provided guidance on methodologies and participated in reviewing and editing the manuscript for clarity and accuracy.

CRediT authorship contribution statement

Lin Zheng: Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Markus Mueller:** Writing – review & editing, Supervision, Methodology. **Chunbo Luo:** Writing – review & editing, Supervision, Methodology. **Xiaoyu Yan:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgement

Lin Zheng would like to acknowledge the financial support for her PhD from the China Scholarship Council and the University of Exeter. We appreciate the support from Dr. Tamaryn Menneer from the University of Exeter, the Smartline Project (England European Regional Development Fund, 05R16P00305 and 05R18P02819), in providing access to original data on the case study properties.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apenergy.2023.122472.

References

[1] Energy Technology Perspectives 2020 – Analysis. IEA; Mar. 19, 2021 [Online]. Available: https://www.iea.org/reports/energy-technology-perspectives-2020.

[2] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev Jan. 2018;81:1192–205. https://doi.org/10.1016/j.rser.2017.04.095.

[3] Petit-Boix A, et al. Application of life cycle thinking towards sustainable cities: a review. J Clean Prod Nov. 2017;166:939–51. https://doi.org/10.1016/j.jclepro.2017.08.030.

[4] Algren M, Fisher W, Landis AE. Chapter 8 - machine learning in life cycle assessment. In: Dunn J, Balaprakash P, editors. Data science applied to sustainability analysis. Elsevier; 2021. p. 167–90. https://doi.org/10.1016/B978-0-12-817976-5.00009-7.

[5] BS EN 15978:2011 - Sustainability of construction works. Assessment of environmental performance of buildings. Calculation method [Online]. Available: https://shop.bsigroup.com/ProductDetail?pid=000000000030256638; Mar 08, 2021.

[6] Basbagill J, Flager F, Lepech M, Fischer M. Application of life-cycle assessment to early stage building design for reduced embodied environmental impacts. Build Environ Feb. 2013;60:81–92. https://doi.org/10.1016/j.buildenv.2012.11.009.

[7] Tollefson J. IPCC says limiting global warming to 1.5 °C will require drastic action. Nature Oct. 2018;562(7726):172–3. https://doi.org/10.1038/d41586-018-06876-2.

[8] de Coninck H, et al. Strengthening and Implementing the Global Response. In: Glob. Warm. 15°C Summ. Policy Mak.; 2018. p. 313–443.

[9] Luo XJ, Oyedele LO. Life cycle optimisation of building retrofitting considering climate change effects. Energ Buildings 2022;258:111830. https://doi.org/10.1016/j.enbuild.2022.111830.

[10] Pryshlakivsky J, Searcy C. Fifteen years of ISO 14040: a review. J Clean Prod Oct. 2013;57:115–23. https://doi.org/10.1016/j.jclepro.2013.05.038.

[11] Buyle M, Braet J, Audenaert A. Life cycle assessment in the construction sector: a review. Renew Sustain Energy Rev Oct. 2013;26:379–88. https://doi.org/10.1016/j.rser.2013.05.001.

[12] Pan W, Li K, Teng Y. Rethinking system boundaries of the life cycle carbon emissions of buildings. Renew Sustain Energy Rev Jul. 2018;90:379–90. https://doi.org/10.1016/j.rser.2018.03.057.

[13] Luo XJ. Retrofitting existing office buildings towards life-cycle net-zero energy and carbon. Sustain Cities Soc Aug. 2022;83:103956. https://doi.org/10.1016/j.scs.2022.103956.

[14] World's fastest building life cycle assessment software - one click LCA. One Click LCA® Software; Oct. 19, 2022 [Online]. Available: https://www.oneclicklca.com/.

[15] IMPACT - BRE Group [Online]. Available: https://bregroup.com/products/impact/; Nov. 09, 2022.

[16] EC3 User guide. Building Transparency; Mar 03, 2023 [Online]. Available, https://www.buildingtransparency.org/ec3-resources/ec3-user-guide/.

[17] tallyCAT beta. Building Transparency; Mar. 03, 2023 [Online]. Available: https://www.buildingtransparency.org/tally/tallycat/.

[18] Atmaca A, Atmaca N. Life cycle energy (LCEA) and carbon dioxide emissions (LCCO2A) assessment of two residential buildings in Gaziantep, Turkey. Energ Buildings Sep. 2015;102:417–31. https://doi.org/10.1016/j.enbuild.2015.06.008.

[19] Peng C. Calculation of a building's life cycle carbon emissions based on Ecotect and building information modeling. J Clean Prod Jan. 2016;112:453–65. https://doi.org/10.1016/j.jclepro.2015.08.078.

[20] Wu X, Peng B, Lin B. A dynamic life cycle carbon emission assessment on green and non-green buildings in China. Energ Buildings Aug. 2017;149:272–81. https://doi.org/10.1016/j.enbuild.2017.05.041.

[21] Zhang X, Liu K, Zhang Z. Life cycle carbon emissions of two residential buildings in China: comparison and uncertainty analysis of different assessment methods. J Clean Prod Sep. 2020;266:122037. https://doi.org/10.1016/j.jclepro.2020.122037.

[22] Xiang-Li L, Zhi-Yong R, Lin D. An investigation on life-cycle energy consumption and carbon emissions of building space heating and cooling systems. Renew Energy Dec. 2015;84:124–9. https://doi.org/10.1016/j.renene.2015.06.024.

[23] Asdrubali F, Baldassarri C, Fthenakis V. Life cycle analysis in the construction sector: guiding the optimization of conventional Italian buildings. Energ Buildings Sep. 2013;64:73–89. https://doi.org/10.1016/j.enbuild.2013.04.018.

[24] Röck M, et al. Embodied GHG emissions of buildings – the hidden challenge for effective climate change mitigation. Appl Energy Jan. 2020;258:114107. https://doi.org/10.1016/j.apenergy.2019.114107.

[25] Kumanayake R, Luo H. A tool for assessing life cycle CO2 emissions of buildings in Sri Lanka. Build Environ Jan. 2018;128:272–86. https://doi.org/10.1016/j.buildenv.2017.11.042.

[26] Sustainability | free full-text | building simplified life cycle CO2 emissions assessment tool (B-SCAT) to support low-carbon building Design in South Korea [Online]. Available: https://www.mdpi.com/2071-1050/8/6/567; Aug 09, 2021.

[27] Roh S, Tae S, Suk SJ, Ford G, Shin S. Development of a building life cycle carbon emissions assessment program (BEGAS 2.0) for Korea's green building index certification system. Renew Sustain Energy Rev Jan. 2016;53:954–65. https://doi.org/10.1016/j.rser.2015.09.048.

[28] D'Amico B, Pomponi F. Accuracy and reliability: a computational tool to minimise steel mass and carbon emissions at early-stage structural design. Energ Buildings Jun. 2018;168:236–50. https://doi.org/10.1016/j.enbuild.2018.03.031.

[29] Malmqvist T, et al. Life cycle assessment in buildings: the ENSLIC simplified method and guidelines. Energy Apr. 2011;36(4):1900–7. https://doi.org/10.1016/j.energy.2010.03.026.

[30] Sousa I, Wallace D, Borland N, Deniz J. A learning surrogate LCA model for integrated product design. 2023.

[31] D'Amico B, et al. Machine learning for sustainable structures: a call for data. Structures Jun. 2019;19:1–4. https://doi.org/10.1016/j.istruc.2018.11.013.

[32] Data Science Applied to Sustainability Analysis - 1st Edition [Online]. Available, https://www.elsevier.com/books/data-science-applied-to-sustainability-analysis/dunn/978-0-12-817976-5; Oct. 11, 2022.

[33] Deng H, Fannon D, Eckelman MJ. Predictive modeling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata. Energ Buildings Mar. 2018;163:34–43. https://doi.org/10.1016/j.enbuild.2017.12.031.

[34] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev Oct. 2009;13(8):1819–35. https://doi.org/10.1016/j.rser.2008.09.033.

[35] Seyedzadeh S, Rahimian FP, Glesk I, Roper M. Machine learning for estimation of building energy consumption and performance: a review. Vis Eng Oct. 2018;6(1):5. https://doi.org/10.1186/s40327-018-0064-7.

[36] Ekici BB, Aksoy UT. Prediction of building energy consumption by using artificial neural networks. Adv Eng Softw May 2009;40(5):356–62. https://doi.org/10.1016/j.advengsoft.2008.05.003.

[37] Veiga RK, Veloso AC, Melo AP, Lamberts R. Application of machine learning to estimate building energy use intensities. Energ Buildings Oct. 2021;249:111219. https://doi.org/10.1016/j.enbuild.2021.111219.

[38] Sharif SA, Hammad A. Developing surrogate ANN for selecting near-optimal building energy renovation methods considering energy consumption, LCC and LCA. J Build Eng Sep. 2019;25:100790. https://doi.org/10.1016/j.jobe.2019.100790.

[39] Li G, Tian W, Zhang H, Fu X. A novel method of creating machine learning-based time series meta-models for building energy analysis. Energ Buildings Feb. 2023;281:112752. https://doi.org/10.1016/j.enbuild.2022.112752.

[40] Chalal ML, Benachir M, White M, Shrahily R. Energy planning and forecasting approaches for supporting physical improvement strategies in the building sector: a review. Renew Sustain Energy Rev Oct. 2016;64:761–76. https://doi.org/10.1016/j.rser.2016.06.040.

[41] Tian Z, Zhang X, Wei S, Du S, Shi X. A review of data-driven building performance analysis and design on big on-site building performance data. J Build Eng Sep. 2021;41:102706. https://doi.org/10.1016/j.jobe.2021.102706.

[42] Qavidel Fard Z, Zomorodian ZS, Korsavi SS. Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. Feb. 2022. https://doi.org/10.1016/j.enbuild.2021.111771.

[43] Zhang Y, O'Neill Z, Dong B, Augenbroe G. Comparisons of inverse modeling approaches for predicting building energy performance. Build Environ Apr. 2015;86:177–90. https://doi.org/10.1016/j.buildenv.2014.12.023.

[44] Menezes AC, Cripps A, Bouchlaghem D, Buswell R. Predicted vs. actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap. Appl Energy Sep. 2012;97:355–64. https://doi.org/10.1016/j.apenergy.2011.11.075.

[45] Carli R, Dotoli M, Pellegrino R, Ranieri L. Using multi-objective optimization for the integrated energy efficiency improvement of a smart city public buildings' portfolio. In: 2015 IEEE International Conference on Automation Science and Engineering (CASE); Aug. 2015. p. 21–6. https://doi.org/10.1109/CoASE.2015.7294035.

[46] Chen Z, et al. A review of data-driven fault detection and diagnostics for building HVAC systems. Appl Energy Jun. 2023;339:121030. https://doi.org/10.1016/j.apenergy.2023.121030.

[47] Balali Y, Chong A, Busch A, O'Keefe S. Energy modelling and control of building heating and cooling systems with data-driven and hybrid models—a review. Renew Sustain Energy Rev Sep. 2023;183:113496. https://doi.org/10.1016/j.rser.2023.113496.

[48] Song Y, Xia M, Chen Q, Chen F. A data-model fusion dispatch strategy for the building energy flexibility based on the digital twin. Appl Energy Feb. 2023;332:120496. https://doi.org/10.1016/j.apenergy.2022.120496.

[49] Pomponi F, Anguita ML, Lange M, D'Amico B, Hart E. Enhancing the practicality of tools to estimate the whole life embodied carbon of building structures via machine learning models. Front Built Environ Apr. 26, 2022;7:2021. https://doi.org/10.3389/fbuil.2021.745598 [Online]. Available:.

[50] Xikai M, Lixiong W, Jiwei L, Xiaoli Q, Tongyao W. Comparison of regression models for estimation of carbon emissions during building's lifecycle using designing factors: a case study of residential buildings in Tianjin, China. Energ Buildings Dec. 2019;204:109519. https://doi.org/10.1016/j.enbuild.2019.109519.

[51] Ye H, et al. Modeling energy-related CO2 emissions from office buildings using general regression neural network. Resour Conserv Recycl Feb. 2018;129:168–74. https://doi.org/10.1016/j.resconrec.2017.10.020.

[52] Płoszaj-Mazurek M, Ryńska E, Grochulska-Salak M. Methods to optimize carbon footprint of buildings in regenerative architectural design with the use of machine learning, convolutional neural network, and parametric design. Energies Jan. 2020;13(20). https://doi.org/10.3390/en13205289. Art. no. 20.

[53] Tsay Y-S, Yeh C-Y, Chen Y-H, Lu M-C, Lin Y-C. A machine learning-based prediction model of LCCO2 for building envelope renovation in Taiwan. Sustainability Jan. 2021;13(15). https://doi.org/10.3390/su13158209. Art. no. 15.

[54] Belyadi H, Haghighat A. Chapter 3 - machine learning workflows and types. In: Belyadi H, Haghighat A, editors. Machine learning guide for oil and gas using Python. Gulf Professional Publishing; 2021. p. 97–123. https://doi.org/10.1016/B978-0-12-821929-4.00001-9.

[55] The-Housing-Stock-of-the-United-Kingdom_Report_BRE-Trust [Online]. Available: https://files.bregroup.com/bretrust/The-Housing-Stock-of-the-United-Kingdom_Report_BRE-Trust.pdf; Apr 03, 2023.

[56] Menneer T, et al. Changes in domestic energy and water usage during the UK COVID-19 lockdown using high-resolution temporal data. Int J Environ Res Public Health Jun. 2021;18(13):6818. https://doi.org/10.3390/ijerph18136818.

[57] Smartline project. Smartline; Sep. 05, 2021 [Online]. Available: https://www.smartline.org.uk.

[58] Williams AJ, Maguire K, Morrissey K, Taylor T, Wyatt K. Social cohesion, mental wellbeing and health-related quality of life among a cohort of social housing residents in Cornwall: a cross sectional study. BMC Public Health Dec. 2020;20(1):985. https://doi.org/10.1186/s12889-020-09078-6.

[59] Zheng L, Mueller M, Luo C, Menneer T, Yan X. Variations in whole-life carbon emissions of similar buildings in proximity: an analysis of 145 residential properties in Cornwall, UK. Energ Buildings Jul. 2023:113387. https://doi.org/10.1016/j.enbuild.2023.113387.

[60] Whole Life Carbon Assessment for the Built Environment, 1st edition. RICS; Sep. 21, 2022 [Online]. Available, https://www.rics.org/uk/upholding-professional-standards/sector-standards/building-surveying/whole-life-carbon-assessment-for-the-built-environment/.

[61] Woods R, Menneer T, Wellaway I, Broughton B, Williams A, Sharpe R, et al. *Smartline Environmental Sensor Data and Utility Usage, 2017–2023.* [data

collection]. UK Data Service. SN 2023;856596. https://doi.org/10.5255/UKDA-SN-856596.

[62] ML | data preprocessing in Python. GeeksforGeeks; Feb 07, 2023 [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/.

[63] LCA Compendium – The Complete World of Life Cycle Assessment. Springer; May 11, 2022 [Online]. Available, https://www.springer.com/series/11776.

[64] Fijten. Translational research on exhaled volatile organic compounds from bedside to bench. Maastricht University; 2017. https://doi.org/10.26481/dis.20171211rf.

[65] Data standardization - an overview | ScienceDirect topics [Online]. Available: https://www.sciencedirect.com/topics/computer-science/data-standardization; Nov 07, 2023.

[66] sklearn.preprocessing.StandardScaler. scikit-learn; Mar. 06, 2023 [Online]. Available, https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[67] Bousquet O, von Luxburg U, Rätsch G. Advanced lectures on machine learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003: revised lectures. In: Lecture notes in computer science, lecture notes in artificial intelligence, no. 3176. Berlin; New York: Springer; 2004.

[68] Principal component analysis: a review and recent developments | philosophical transactions of the Royal Society a: mathematical, physical and engineering sciences. May 22, 2023. https://doi.org/10.1098/rsta.2015.0202 [Online]. Available:.

[69] Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg May 2018;126(5):1763. https://doi.org/10.1213/ANE.0000000000002864.

[70] Gu J, et al. Selection of key ambient particulate variables for epidemiological studies — applying cluster and heatmap analyses as tools for data reduction. Sci Total Environ Oct. 2012;435–436:541–50. https://doi.org/10.1016/j.scitotenv.2012.07.040.

[71] JetBrains DataSpell: the IDE for data scientists. JetBrains; Nov 02, 2022 [Online]. Available: https://www.jetbrains.com/dataspell/.

[72] Python release Python 3.9.0. Python.org; Jun 09, 2022 [Online]. Available: https://www.python.org/downloads/release/python-390/.

[73] Scikit-learn: machine learning in Python — scikit-learn 1.1.3 documentation [Online]. Available: https://scikit-learn.org/stable/; Nov. 02, 2022.

[74] Package overview — pandas 1.5.1 documentation [Online]. Available: https://pandas.pydata.org/docs/getting_started/overview.html; Nov. 02, 2022.

[75] Waskom M. seaborn: statistical data visualization. J Open Source Softw Apr. 2021;6(60):3021. https://doi.org/10.21105/joss.03021.

[76] Yan X, Su X. Linear regression analysis: Theory and computing. Singapore; Hackensack, NJ: World Scientific; 2009.

[77] Freedman D. Statistical models: Theory and practice. Cambridge: Cambridge University Press; 2005. https://doi.org/10.1017/CBO9781139165495.

[78] Seal HL. Studies in the history of probability and statistics. XV the historical development of the gauss linear model. Biometrika Jun. 1967;54(1–2):1–24. https://doi.org/10.1093/biomet/54.1-2.1.

[79] Bedoui A, Lazar NA. Bayesian empirical likelihood for ridge and lasso regressions. Comput Stat Data Anal May 2020;145:106917. https://doi.org/10.1016/j.csda.2020.106917.

[80] Zwillinger D. Standard MathematicAL TABLES and formulae. 2003. p. 840.

[81] Encyclopedia of Statistical Sciences. Vol. 1. John Wiley & Sons; 2005.

[82] Stephanie. Lasso regression: simple definition', statistics how to [Online]. Available: https://www.statisticshowto.com/lasso-regression/; Oct. 31, 2022.

[83] Emami Javanmard M, Ghaderi SF, Hoseinzadeh M. Data mining with 12 machine learning algorithms for predict costs and carbon dioxide emission in integrated energy-water optimization model in buildings. Energ Conver Manage Jun. 2021; 238:114153. https://doi.org/10.1016/j.enconman.2021.114153.

[84] Predicting building-related carbon emissions: a test of machine learning models | SpringerLink. Jan. 23, 2023. https://doi.org/10.1007/978-3-030-52067-0_11 [Online]. Available.

[85] Decision tree regression [Online]. Available: http://www.saedsayad.com/decision_tree_reg.htm; Jan. 23, 2023.

[86] Wu D. Supplier selection: a hybrid model using DEA, decision tree and neural network. Expert Syst Appl Jul. 2009;36(5):9105–12. https://doi.org/10.1016/j.eswa.2008.12.039.

[87] Predicting disease risks from highly imbalanced data using random forest | SpringerLink. May 22, 2023. https://doi.org/10.1186/1472-6947-11-51 [Online]. Available.

[88] Breiman L. Random forests. Mach Learn Oct. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[89] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal Feb. 2002;38 (4):367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

[90] Awad M, Khanna R. Support vector regression. In: Awad M, Khanna R, editors. Efficient learning machines: Theories, concepts, and applications for engineers and system designers. Berkeley, CA: Apress; 2015. p. 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4.

[91] You H, et al. Comparison of ANN (MLP), ANFIS, SVM, and RF models for the online classification of heating value of burning municipal solid waste in circulating fluidized bed incinerators. Waste Manag Oct. 2017;68:186–97. https://doi.org/10.1016/j.wasman.2017.03.044.

[92] Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. J Math Psychol Aug. 2018;85: 1–16. https://doi.org/10.1016/j.jmp.2018.03.001.

[93] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. In adaptive computation and machine learning. Cambridge, Mass: MIT Press; 2006.

[94] Beckers T. An introduction to Gaussian process models. arXiv preprint arXiv: 2102.05497. Feb. 10, 2021 Accessed: May 22, 2023. [Online]. Available: http://arxiv.org/abs/2102.05497.

[95] Antanasijević D, Pocajt V, Ristić M, Perić-Grujić A. Modeling of energy consumption and related GHG (greenhouse gas) intensity and emissions in Europe using general regression neural networks. Energy May 2015;84:816–24. https://doi.org/10.1016/j.energy.2015.03.060.

[96] Ding L, Rangaraju P, Poursaee A. Application of generalized regression neural network method for corrosion modeling of steel embedded in soil. Soils Found Apr. 2019;59(2):474–83. https://doi.org/10.1016/j.sandf.2018.12.016.

[97] Specht DF. A general regression neural network. IEEE Trans Neural Netw Nov. 1991;2(6):568–76. https://doi.org/10.1109/72.97934.

[98] '3.1. Cross-validation: evaluating estimator performance', scikit-learn [Online]. Available, https://scikit-learn/stable/modules/cross_validation.html; Nov. 09, 2022.

[99] Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput Apr. 2011;21(2):137–46. https://doi.org/10.1007/s11222-009-9153-8.

[100] 'An introduction to statistical learning', an introduction to statistical learning [Online]. Available: https://www.statlearning.com; Nov. 09, 2022.

[101] Minewiskan. Scatter plot (analysis services - data mining) [Online]. Available: https://learn.microsoft.com/en-us/analysis-services/data-mining/scatter-plot-analysis-services-data-mining?view=asallproducts-allversions; Oct 20, 2023.

[102] Adetunji AB, Akande ON, Ajala FA, Oyewo O, Akande YF, Oluwadara G. House Price prediction using random forest machine learning technique. Procedia Comput Sci Jan. 2022;199:806–13. https://doi.org/10.1016/j.procs.2022.01.100.

[103] Liitiäinen E, Verleysen M, Corona F, Lendasse A. Residual variance estimation in machine learning. Neurocomputing Oct. 2009;72(16):3692–703. https://doi.org/10.1016/j.neucom.2009.07.004.

[104] Dangeti P. Statistics for machine learning. Packt Publishing Ltd; 2017.

[105] Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA. A survey of forecast error measures. 2013. p. 7.

[106] Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal Jan. 2008;52(4):2249–60. https://doi.org/10.1016/j.csda.2007.08.015.