Expert Systems With Applications 225 (2023) 120109

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Building energy performance prediction: A reliability analysis and evaluation of feature selection methods

Razak Olu-Ajayi, Hafiz Alaka [*], Ismail Sulaimon, Habeeb Balogun, Godoyon Wusu, Wasiu Yusuf, Muideen Adegoke

Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield AL10 9AB, UK

ARTICLE INFO

Keywords:
Feature selection
Building energy performance
Energy efficiency
Machine learning
Energy prediction

ABSTRACT

The advancement of smart meters using evolving technologies such as the Internet of Things (IoT) is producing more data for the training of energy prediction models. Since many machine learning techniques were not premeditated to handle a large number of irrelevant features, it has engendered the search for optimal techniques to decrease the generated features and potentially identify the most relevant features that have an impact on building energy efficiency. Feature selection is considered one of the most suitable methods of pinpointing the best features combination. However, the fraction of studies that deliver comprehensive insights on the incorporation of feature selection with machine learning is still limited, notwithstanding the capabilities of feature selection to produce a good result in terms of accuracy and speed. To address this gap, this study investigates feature selection methods centred on building energy consumption prediction using machine learning. This study conducted a comparative analysis of 14 machine learning algorithms on 5 different data sizes and explored the effect of 7 feature selection methods on model performance for predicting energy consumption in buildings. Furthermore, this study identifies the most effective feature selection methods and machine learning models for energy use prediction. The experimental results demonstrate that feature selection can affect model's performance positively or negatively, depending on the algorithm employed. Nevertheless, the filter method was noted as the most appropriate method for most Machine Learning (ML) classification algorithms. Moreover, Gradient Boosting (GB) was identified as the most effective model for predicting energy performance in buildings. Additionally, the reliability analysis confirms the hypothesis that "the larger the data, the more accurate the result" but only for specific algorithms such as Deep Neural Networks (DNN). This study also presents the theoretical and practical implications of this research.

1. Introduction

Buildings are major energy consumers that significantly contribute to vital energy-related environmental issues such as climate change and air pollution, among others (Allouhi et al., 2015; Dandotiya, 2020; Wang et al., 2005). The rate of building energy consumption is increasing considerably, representing around 30% of global energy usage, with the prospect of rapid growth in the nearest future (Amasyali & El-Gohary, 2018; Dong et al., 2021; Zhong et al., 2019). For example, the United Kingdom (UK) buildings in particular, consume over 29% of the overall energy consumption ("BEIS", 2019; Building Energy Efficiency Survey, 2016). However, it is presumed that accurate prediction of building energy use is the most effective method of understanding building energy efficiency (Fathi et al., 2020; Lei et al., 2021; Hai-xiang Zhao & Magoulès, 2012a). Therefore, based on the recognition of Machine Learning (ML) for the generation of good performance in prediction tasks (Adegoke, Hafiz, Ajayi, & Olu-Ajayi, 2022; Zhou & Chen, 2021; Canales, 2016; Olu-Ajayi and Alaka, 2021; Vorobeychik and Wallrabenstein, 2013). Researchers have proposed several Machine Learning (ML) algorithms for developing energy predictive models, that potentially produced good results (Bourhnane et al., 2020; Castelli et al., 2015; Lei et al., 2021; Qiong Li et al., 2010).

Despite the good predictive performance of ML algorithms, none has been identified as the best for accurate energy predictions. One of the most broadly unaddressed issues in energy consumption literature, which affects ML algorithms performance is feature or variable selection (Hsu, 2015). Identifying the most relevant features strongly affects the

* Corresponding author.
E-mail addresses: r.olu-ajayi@herts.ac.uk (R. Olu-Ajayi), h.alaka@herts.ac.uk (H. Alaka), i.sulaimon@herts.ac.uk (I. Sulaimon), h.balogun2@herts.ac.uk (H. Balogun), g.e.wusu@herts.ac.uk (G. Wusu), w.a.yusuf@herts.ac.uk (W. Yusuf), m.adegoke@herts.ac.uk (M. Adegoke).

https://doi.org/10.1016/j.eswa.2023.120109
Received 6 January 2023; Received in revised form 21 March 2023; Accepted 8 April 2023
Available online 15 April 2023
0957-4174/

Nomenclature

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

accuracy of the predictive model (Alaka et al., 2018; Pirbazari et al., 2019). Several studies often select variables derived from academic literature to develop ML models. However, before model development, the application of feature selection method using statistical techniques is considered one of the most effective methods for identifying the most important variables or features (Hai-xiang Zhao & Magoulès, 2012a). Feature Selection (FS) aims to improve the performance of ML models by eliminating the unimportant and irrelevant noisy features, thus improving the quality of the dataset (Asir et al., 2016). These redundant features are removed to reduce the input dimensionality. Feature selection decreases the computational time as well as the difficulty of training and testing a classification model. Therefore, it engenders more cost effective predictive models (Effrosynidis & Arampatzis, 2021). Furthermore, selecting the most relevant features simplifies the calculation and reduces the dimensionality (HaiXiang Zhao & Magoulès, 2012). Although there are several advantages associated with the application of feature selection, their drawbacks include the unfavorability of some feature selection methods on ML algorithms (Balogun et al., 2021).

Although features (or input variables) are mostly selected based on domain knowledge in the field of building energy use prediction, [e.g., (Bagnasco et al., 2015; Bourhnane et al., 2020; Ding & Liu, 2020; Dong et al., 2021; K. Li et al., 2018)]. Some studies have applied the feature selection method in the development of energy predictive models (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b). However, the fraction of studies that deliver comprehensive insights on the incorporation of feature selection with machine learning is still limited, notwithstanding the capabilities of feature selection to produce a good result in terms of accuracy and speed. There are insufficient studies that have compared and analysed various feature selection techniques in predicting building energy performance. To address this gap, this study aims to identify the most suitable feature selection method for for building developing a ML model for energy consumption prediction. Additionally, it is a general hypothesis in the machine learning world, that sample size has a positive impact on the predictive performance of a machine learning model [i.e. the larger the data, the more accurate the result (Goyal et al., 2020; Kabir, 2020; Kaur & Gupta, 2017; Lee et al., 2011; Olu-Ajayi et al., 2022a)]. Therefore, this study will also investigate the effect of sample size on the performance of ML supervised learning models.

This study will develop building energy usage prediction models using various ML classification algorithms [such as Random Forest (RF) (Carrera et al., 2021; Chen et al., 2019; C. Li et al., 2018; Pham et al., 2020), Support Vector Machine (SVM) (Dong et al., 2005; Jing et al., 2022, p. 202; Liu et al., 2020; Shao et al., 2020; Zhong et al., 2019) etc.], and applying various feature selection methods [such as random forest (Z. Dong et al., 2021; Z. Wang et al., 2018; Zhang & Wen, 2019b) and chi square (Bahassine et al., 2020; Sumaiya Thaseen & Aswani Kumar, 2017) etc.] and different sample sizes. This research aims to conduct an unbiased comparison of feature selection methods to determine the most effective FS method and ML algorithm for building energy use prediction. The main objectives used to achieve the aim of this research are enumerated below:

- To explore the suitability or benefits of feature selection methods on various ML classification algorithms.
- To investigate the effect of sample size on the performance of ML classification models.
- To identify the most suitable feature selection technique for buildingsbuilding energy performance prediction.

The rest of this paper is structured as follows: Section 2 presents a review of feature selection. Section 3 elucidates the data gathered, methodology of research, the method of pre-processing data, model development and evaluation measures. Section 4 discusses the performance results and findings. Section 5 presents the theoretical and practical implications. Section 6 delivers the overall conclusion and future recommendations.

2. Literature review

Researchers have established that the appropriate choice of features or variables is closely connected to the increase in performance accuracy of a model (Zhao & Magoulès, 2012; Zhang & Wen, 2019a). Feature selection is recognised as a data pre-processing technique for efficient data preparation (mainly high-dimensional data) in machine learning problems (Asir et al., 2016; Li et al., 2017; Maldonado & Weber, 2009). High dimensional features often incur a high computational cost, while low dimensional data decreases the probability of overfitting. Feature selection measures the relevance dependency of each feature with the output label. It also identifies and eliminates the irrelevant and redundant features in a dataset (Chandrashekar & Sahin, 2014). Irrelevant features are features that have no impact on the target function in any way, while redundant features are features that add nothing to the target function (Dash & Liu, 1997). The elimination of irrelevant and redundant variables often reduces the data, leading to enhancement in the classification performance.

Machine learning algorithms can often perform classification based on a set of features, and feature selection is essential in classification (Blum & Langley, 1997; Maldonado & Weber, 2009). Generally, a feature is a singular quantifiable property of a process being perceived (Chandrashekar & Sahin, 2014). In real-world situations, data is often represented using numerous features. However, in most cases, only a few of these features may be related to the target output (Kira & Rendell, 1992). These unrelated features that constitute no correlation to labels serve as pure noise, which could lead to bias in prediction, thereby diminishing the classification performance (Kunasekaran & Sugumaran, 2016). Such situations require feature selection to speed up the learning process and enhance the quality of data.

The integration of feature selection methods with machine learning for predicting building energy usage has slowly been explored more recently. For example, random forest and pearson's correlation coefficient were applied for ranking a total of 124 features for building energy use prediction (Zhang et al., 2018). Also, (Faisal et al., 2019) utilized recursive feature elimination and mutual information methods to calculate the importance of the input features for predicting electricity consumption. It was concluded that the results produced using the selected features outperformed the results using the original features. Furthermore, it is suggested that feature selection can aid the reduction of frequently experienced overfitting issue.

Additionally, (Zhang & Wen, 2019a) conducted a novel exploration of feature selection methods. Initially, the original feature sets comprised of 278 features. However, using domain knowledge 22 features were chosen, subsequently using Pearson correlation coefficient 29 features were selected. Lastly, multivariate adaptive regression splines was employed for thorough selection which elicited 14 features. (Paudel et al., 2017) employed feature selection in the prediction of energy consumption and emphasized the benefits of feature selection for increase in accuracy levels and decrease in computational times.

In this research, given that the building energy dataset comprises of building properties, feature selection using machine learning algorithms has a great possibility of identifying the most relevant features. Three key feature selection methods have been established for feature selection namely filter, wrapper and embedded methods (Maldonado & Weber, 2009; Zhang & Wen, 2019b). Filter methods select features centred on performance measures without consideration of the type of modelling algorithm utilized (Jović et al., 2015). Few studies have applied filter techniques to select relevant features in building energy use prediction. For instance, Kapetanakis et al. (2017) used linear correlation feature selection technique in the development of a thermal load prediction model. It was concluded that the accuracy remained on the same level with the FS application. Kusiak et al. (2010) also applied boosting tree feature selection technique in the selection of the relevant variables for predicting building steam load. The relevant input variables included maximum, minimum and mean of outdoor air temperature (Kusiak et al., 2010).

Filter methods provide more generality as they are independent of the chosen algorithm. They are fast in execution and computationally inexpensive in comparison to the wrapper and embedded techniques (Asir et al., 2016). The wrapper and embedded methods are considered less efficient than filter methods for high-dimensional data processing (Iqbal et al., 2020). The filter method is implemented based on the common characteristics of features such as distance between classes and statistical dependencies (i.e., each feature is allotted a statistical score). However, this method is blind to any connections between the features. Thus, this method will not recognize features that can be relevant when combined other features (Bommert et al., 2020). This method removes irrelevant features reducing the feature set dimensionality without losing much model accuracy (Zhang & Wen, 2019b). There are different types of filter methods namely chi square, boosting tree, linear correlation, and ANOVA, among others.

Filter methods utilize methods of variable or feature ranking as the customary criteria for feature selection (Aziz et al., 2017). A group of statistical methods are employed to grade each feature or the whole feature sets, Contingent on if multiple features can be assessed simultaneously. Contrasting to filter methods that implement feature selection independently of the development of the prediction model, the wrapper method utilizes a ML algorithm for feature subset evaluation with regard to classification error and accuracy.

The major difference between the wrapper and filter is the evaluation criteria. Kohavi and John (1997) developed the wrapper-based feature selection technique for selecting relevant features (or input variables) from the dataset. The performance of this technique is often evaluated based on classification accuracy using naïve bayes and decision tree classifiers. However, the wrapper method has difficulties e.g. overfitting, overhead searching and prolonged computational time (Kohavi & John, 1997). Furthermore, the wrapper method utilizes the machine learning algorithm to assess the produced subsets by using the searching technique, making it more computationally complex. Thus, these techniques are not appropriate for high-dimensional space (Asir et al., 2016). There are various types of wrapper methods such as recursive feature elimination, forward feature selection, random forest and boruta among others.

Various studies have adopted the wrapper method in the building energy use prediction field. Fan et al., 2014 employed recursive feature elimination to conduct feature selection for eight machine learning algorithms in predicting next day building energy consumption. After the evaluation of the algorithms, it was observed that Random Forest (RF) and Support Vector Regression (SVR) produced the best result in terms of model performance (Fan et al., 2014). Also, Ahmad et al., 2017a utilized random forest filter techniques in the development of Random Forest (RF) and Artificial Neural Networks (ANN) for the building energy use prediction. The filter method produced a good performance in ANN than RF. Likewise, Dong et al. (2021) applied the RF feature selection method and employed stacking, ANN and Support Vector Regression (SVR) for forecasting hourly energy use. Stacking emerged as the best among other algorithms (Dong et al., 2021). Furthermore, Kolter and Ferreira (2011) implemented a forward selection method in predicting energy consumption in a building. The selection of best performance was centred on Root Mean Squared Error (RMSE) of a predictive model. It was concluded that the RMSE of an energy predictive model can be decreased by utilizing the selected features.

Table 1
Merits and demerits of feature selection.

Method	Merits	Demerits
Filter	• Independent of the selected algorithm. • Fast execution • Low computational cost	• Ignores connections between the features • Ignores interaction with the classifier
Wrapper	• Simple • Considers feature dependencies	• Chances of overfitting • More computationally complex • Long computational time
Embedded	• Fast execution • lower chance to overfitting.	• Limited to specific algorithms and hence they are not always used

The embedded method assesses the usefulness of features similar to the wrapper method (HaiXiang Zhao & Magoulès, 2012). However, this method implements feature selection during the algorithm's execution; thus, these methods are embedded normally or as an extended functionality in each regression or classification algorithm. The popular embedded methods include some decision tree algorithms such as Embedded Random Forest, Classification and Regression Tree (CART), among others. Furthermore, embedded methods can apply feature selection by the algorithms training process. Thus, due to the abstention of retraining of the specific algorithms, it suffers less computational burden. However, it is peculiar to specific algorithms and hence they are not always used (Zhang et al., 2019).

Embedded methods employ the fundamental characteristic of ML algorithms to execute feature selection (Ang et al., 2016). Embedded methods have different methods: during the process of training the model, features with smaller correlation coefficient values are eliminated recursively through the use of a support vector machine. Another method is application of feature selection as an embedded function during the training process (Table 1).

3. Data and methodology

This research analyses the effect of feature selection methods on various ML algorithms in the prediction of energy use in buildings. This research will further substantiate or invalidate the suitability of feature selection methods for some or all ML classification algorithms. Hence, the question "Is there a feature selection technique considered suitable for all machine learning classification algorithms?". Additionally, based on the hypothesis, the greater the data, the better the model's performance (Dalal, 2018; Goyal et al., 2020; Kabir, 2020; Lee et al., 2011; Olu-Ajayi et al., 2022a), this study employs different data sizes in developing all ML models. This section will describe the different datasets and techniques (feature selection methods and ML classification algorithms) utilized in this study. The schematic diagram of this research is displayed in Fig. 1 below. This section consists of major processes namely data gathering, data pre-processing, feature selection, correlation analysis, model training and model testing.

3.1. Data collection

This study utilized three types of datasets for classification namely building metadata, meteorological and energy dataset. These datasets contain several features considered pertinent to energy performance prediction which has been employed in various studies (Feng & Zhang, 2020; Olu-Ajayi & Alaka, 2021; Wang et al., 2021). The identification of the most relevant features that generate good predictions will thereby improve the classification accuracy for subsequent predictions of the annual energy performance of buildings. These datasets were collected

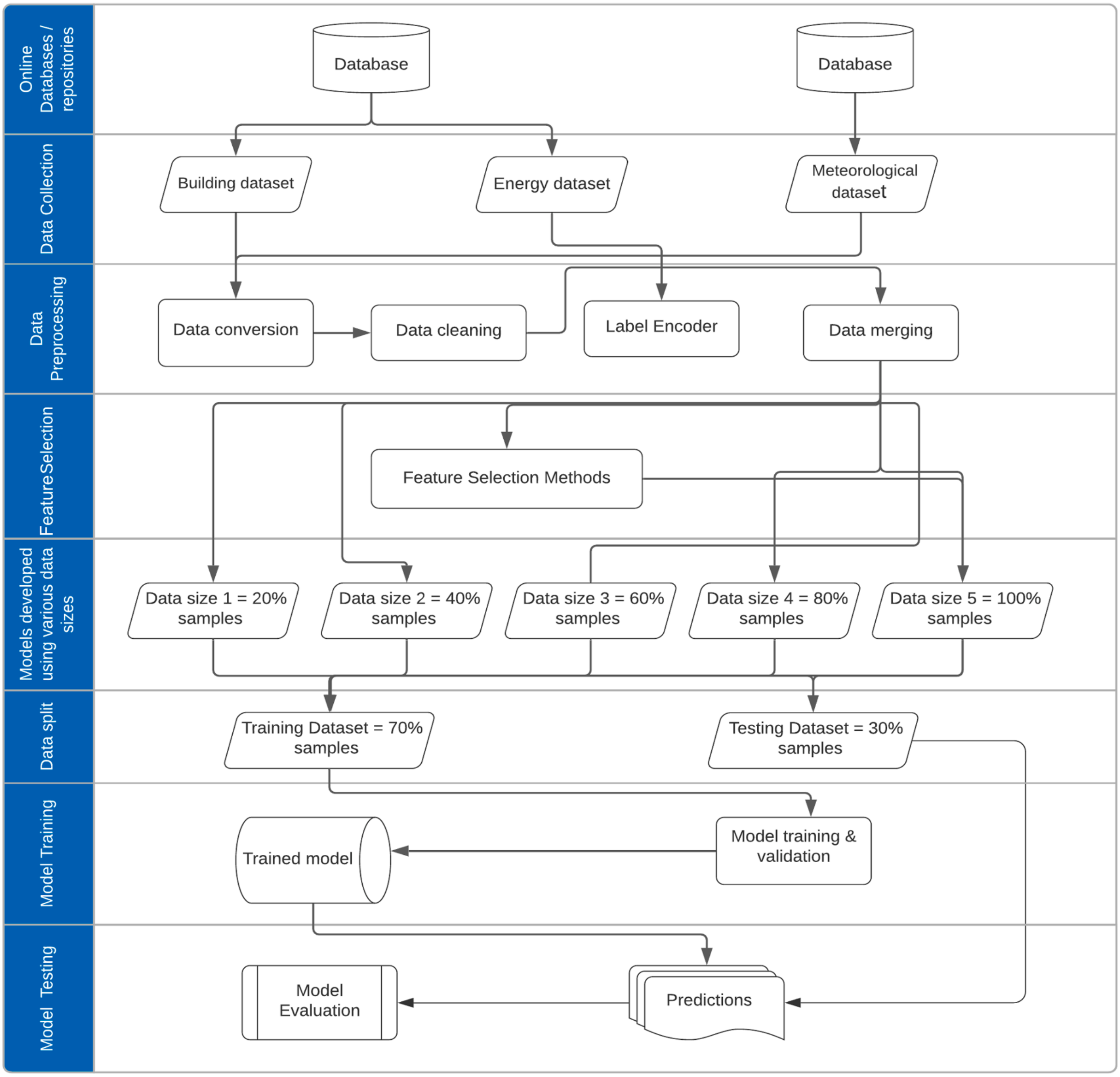

Fig. 1. The framework of this Research.

for many residential buildings in the United Kingdom (UK). This study uses only residential building data as they are the majority of building stock in the UK.

Building Dataset: The building metadata was gathered from the Ministry of Housing Communities and Local Government (MHCLG) repository. The dataset consisted of the properties of each building. The building and energy datasets were collected for 60,000 various types of residential buildings. Fig. 2 shows the proportion for each type of building in the dataset. The types of buildings include flats, bungalows, maisonettes, and houses.

The 60,000 residential buildings are situated within six area postcodes in the United Kingdom namely Hartlepool, Middlesbrough, Redcar and Cleveland, Darlington, Halton, and Warrington. Building and energy related datasets were collected for 10,000 buildings for each area postcode. The building consists of various features that are readily available at the conceptual stage of buildings. Several studies have considered features such as floor, window, and wall type as important input variable (Marino et al., 2017; Marwan, 2020; Tahmasebi et al., 2011). However, this study will further identify the most relevant parameters required for predicting the annual energy performance of buildings. The features selected include some common features such as

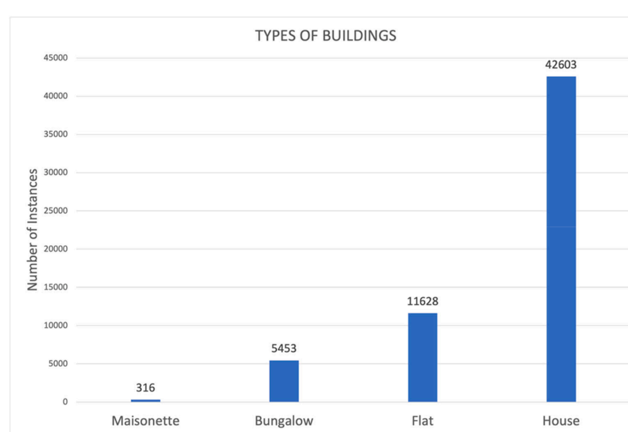

Fig. 2. Proportion of Buildings.

floor area, floor level, and window description among others, which were considered as the independent variable in other studies (Goyal et al., 2020; Kabir, 2020; Lee et al., 2011; Olu-Ajayi et al., 2022a). Fig. 3 displays the independent variables from the building dataset, weather related dataset and the annual energy rating which was employed as the dependent variable.

As displayed in Fig. 3, other features such as window energy efficiency, and window environmental efficiency among others. This is the rate of the energy efficient properties of the building such as window energy efficiency (double glazing, triple glazing), wall energy efficiency (wall types that decreases the air inflation and restrict heat flow) is graded very good to very poor. Likewise, the environmental efficiency is concerned with the rating of the quality of building properties in terms of environmental friendliness.

Weather-related dataset: In this study, the employed meteorological or weather dataset was collected from the Meteostat database. The meteorological dataset is considered one of the key variables in energy use prediction (Ding & Liu, 2020). The scale of the weather-related data gathered was daily from 1st January 2020 till 31st December 2020. These data were averaged to calculate the annual meteorological data for the year 2020, relative to each building's data collected. This weather-related data was gathered using each building's area postcode. The weather-related data was gathered for six areas namely Hartlepool, Middlesbrough, Redcar and Cleveland, Darlington, Hilton and Warrington. The weather-related data comprised of wind speed, pressure, and temperature as shown in Fig. 3 above. The reason for the selection of buildings from different area postcodes is to ensure an unbiased classification. Fig. 4 shows the monthly weather temperature of the five area postcodes.

Energy rating dataset: The energy rating data was also amassed from the Ministry of Housing Communities and Local Government (MHCLG) repository. The energy rating of each building for the year 2020 consist of both high and low energy grade. The building energy rating is based on UK standard rating scale issued in the form of Energy Performance Certificate (EPC) to alert building landlords of present energy rating, energy cost and effectual recommendations on improving energy efficiency and saving money (Curtis et al., 2014). The EPC energy efficiency rating was utilized in the development of a classification model as the target variable. The energy efficiency rating ranges from A to G with 'A' denoted as the most energy efficient and 'G' as the least energy efficient (i.e., 92+ = 'A', 81–91 = 'B', 69–80 = 'C', 55–68 = 'D', 39–54 = 'E', 21–38 = 'F', 01–20 = 'G'). The proportion of the energy rating for each building is displayed in Fig. 5. To avoid the bias of data categories imbalance, multiple evaluation measures were employed (i.e., F1 which computes the mean of precision and recall. Precision been the number of selected instances that are relevant, recall the number of relevant instances that are selected).

3.2. Data pre-processing

Data pre-processing has significant impacts on the machine learning algorithms performance (Kotsiantis et al., 2006). The objective of data pre-processing is to handle raw data imperfections and irregularities such as high dimensionality, noise, missing data, outliers, inconsistencies and imbalanced data (Benhar et al., 2020). Although data pre-processing is computationally expensive and time consuming, it is required to ensure quality assurance of the database and to avoid difficulties during model development (Shapi et al., 2021). Without the implementation of adequate pre-processing of data, several complexities could emerge that affect the model performance such as missing or abnormal values, noise etc. Missing data means the absence of one or more entries in the matrix containing the experimental dataset (Mishra et al., 2020). The missing data identified in the building and weather-related dataset were handled in different ways. Newgard and Lewis (2015) proposed the mean value imputation for handling missing data which was utilized in the weather-related dataset. The building instances containing missing data were removed. The total number of instances removed from the building dataset is 9,461. The building dataset contains categorical features such as the roof energy efficiency and windows energy efficiency among others as shown in Fig. 3. The

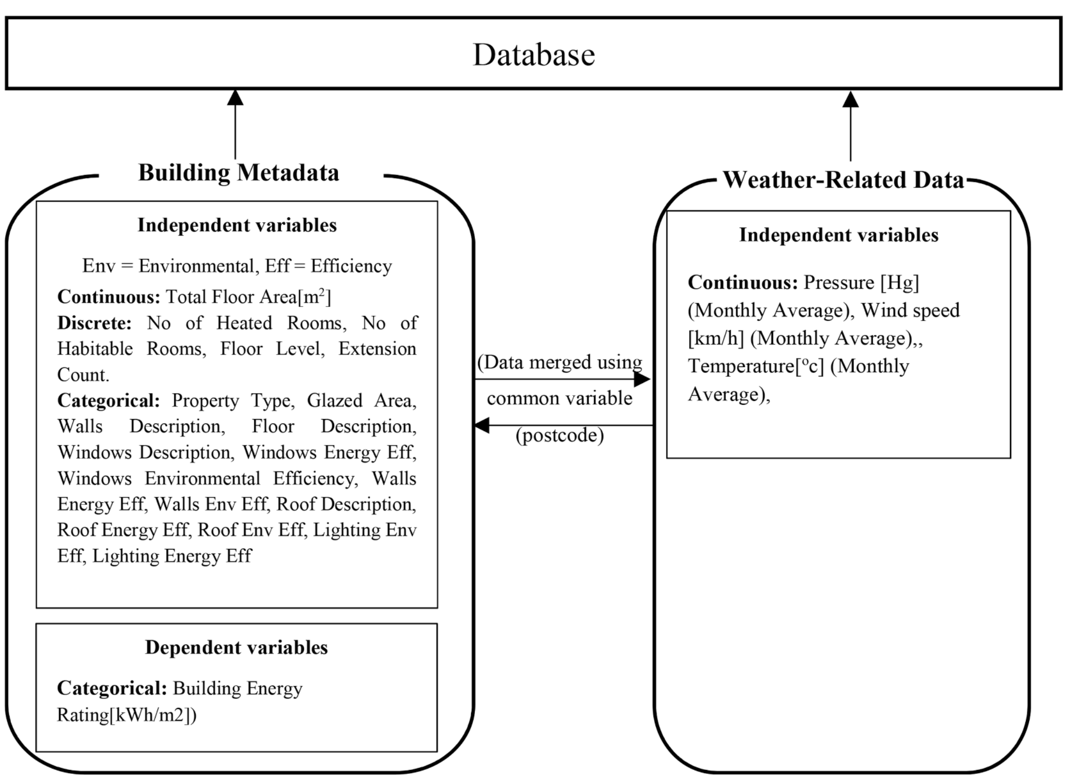

Fig. 3. Building and Weather-related variables collected.

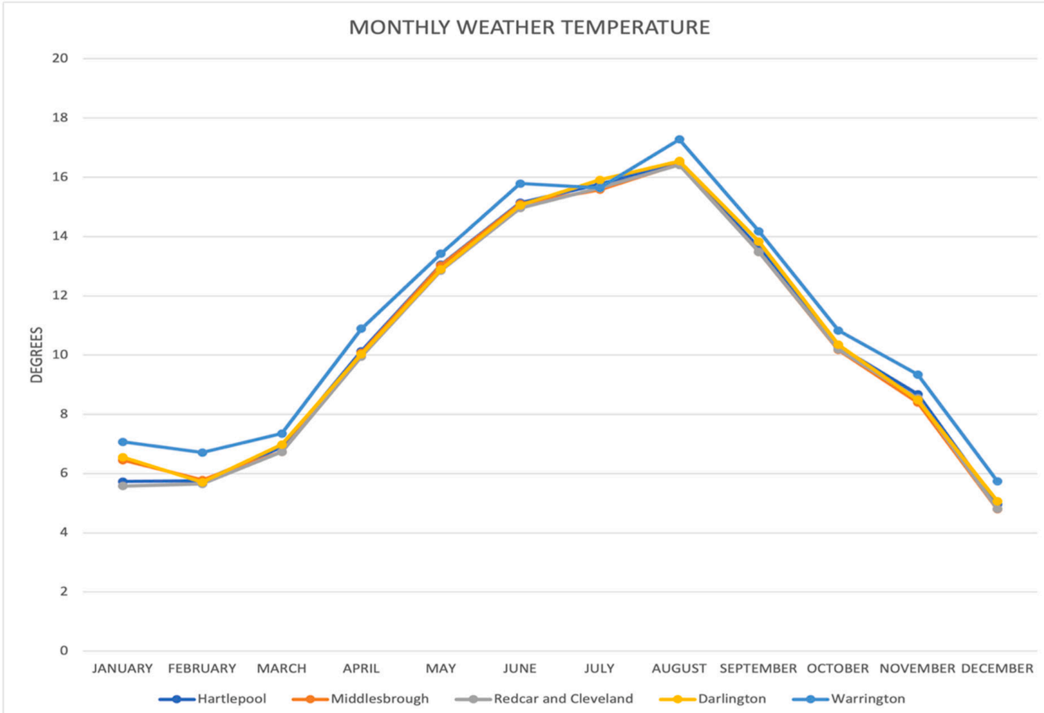

Fig. 4. Monthly weather temperature.

categorical data were allotted values, (e.g., very good = maximum value (5) and very poor = minimum value (0)) to transform the data into an appropriate format for the ML algorithm.

Label encoder: The label encoder is used to encode or convert labels to numeric format the machine can process. For example, converting a variable list called morality containing good and bad to 1 and 2 respectively. The label encoder of the sklearn python package was utilized to convert the labels of the energy rating dataset.

Data Merging: This is the process of consolidating various datasets. The building and weather-related datasets were amalgamated using each location postcode to match all building instances with their corresponding weather-related data. These datasets (building and weather-related datasets) were merged using the panda python package. The merged data constituted a total of 1,320,000 data points.

Data Size: To conduct the investigation of the effect of data size on each ML model's performance, the data utilized for training each model was fixed to 20%, 40%, 60%, 80% and 100% for clear comparison as shown in Fig. 1 above. Therefore, the dissimilar quantities of training sets signify different types of data, for example, 20% and 40% represents the inadequate amount of data, and 60% represent good while 80% and 100% represent an adequate amount of data.

3.3. Feature selection methods

This study utilized seven techniques for the selection of the most relevant features for energy use prediction. These methods include three filter, two wrapper and two embedded methods.

3.3.1. Filters

Chi-square: This is a type of univariate filter FS test that calculates the deviation from the projected distribution considering the feature occurrence is independent of the label values (Sumaiya Thaseen & Aswani Kumar, 2017). Like any univariate method, chi square is calculated between each feature and target or dependent feature, and then the presence of a correlation between them is detected. Subsequently, a low score is assigned if the target variable is independent of the feature while, if the target variable is dependent on the feature, the feature is considered is important (Effrosynidis & Arampatzis, 2021). Therefore, the higher the Chi-Square value means the more relevant the feature.

Mutual Information: This is a type of filter FS method proposed by (Battiti, 1994). It is also known as information gain. Mutual information aims to amplify the relevance between the input and output features, and decrease the redundancy of the chosen features (Amiri et al., 2011). Subsequently, if the information gain of a feature is high, it is considered relevant. However, mutual information does not identify redundant features, because the features are chosen in a univariate way (Effrosynidis & Arampatzis, 2021).

ANOVA: This is a type of univariate filter-based technique that utilizes variance to detect the separability of each feature between classes (Ding et al., 2014).

3.3.2. Wrappers

Permutation importance: This is a heuristic wrapper method centred on repeated permutations of the resulting vector for forecasting the distribution of calculated importance for each feature in a non-informative way (Effrosynidis & Arampatzis, 2021).

Recursive Feature Elimination: This is a type of multivariate wrapper technique that utilizes THE decision tree classifier for training the model repeatedly using the existing features. The least important features are then eliminated using the weight of the algorithm as a ranking measure (Seijo-Pardo et al., 2019).

3.3.3. Embedded

Embedded Random Forest: This is an embedded method using the random forest algorithm. The significance of each feature is calculated by conducting random permutations of features in the out-of-bag set and measuring the increase in misclassification level in comparison to the out-of-bag set in default state (Effrosynidis & Arampatzis, 2021).

ExtraTreesClassifier: ExtraTrees abbreviated as extremely randomized trees, is a form of the random forests which performs randomization at every step for the selection of an optimal split. In contrast to random forests where the split features are centred on a grade, ExtraTrees implements a split measure of the random and

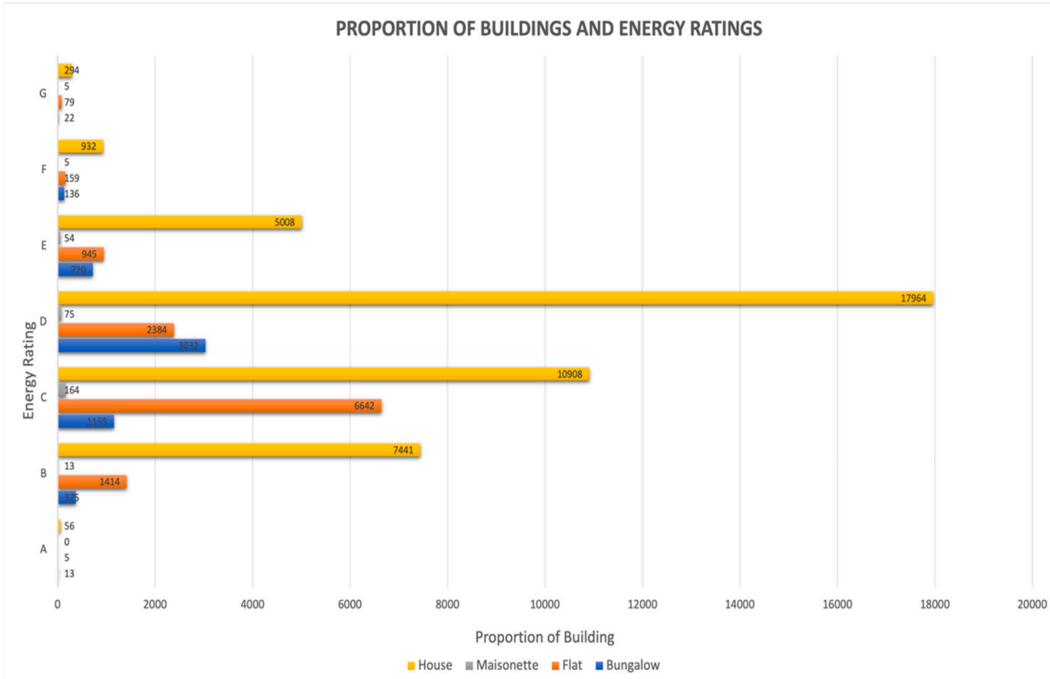

Fig. 5. The proportion of buildings with energy rating.

considers the whole training set (Sharma et al., 2019).

Following the application of the feature selection methods, the 10 highest ranked features were selected, shown in Table 2 above. Amongst these, the most common highest ranked features among the feature selection method are walls description, floor description, walls energy efficiency, roof description. These are important features of a building and have an effect on the energy use of the building. For instance, walls are one of the most essential features of a building, as the selection of the most appropriate type of walls has an effect on the energy performance of a building (i.e., the thickness of the wall will reduce the use of heating in the winter season) (Marwan, 2020).

3.4. Correlation

After feature selection, some variables such as window environmental efficiency, wind speed etc were identified to have comparatively lower values of importance as shown in Table 2. Nonetheless, low importance variables do not directly mean the variable is of low relevance or irrelevant to the target variable. Thus, some variables were significantly correlated with each other and the result for the evaluation of this correlation is provided in Fig. 6 below. The correlation matrix plot shows the correlation between features (building and meteorological variables) and target (energy efficiency rating). As shown in Fig. 6, the diagonal line showing one represents the correlation of the features

Table 2
Rank of each feature selected using various feature selection methods.

Features	Chi-square	Mutual Information	ANOVA	Permutation importance	Recursive Feature Elimination	Embedded Random Forest	ExtraTrees Classifier
Total Floor Area[m✓]	✓			✓	✓	✓	✓
Property Type		✓		✓	✓		✓
Extension Count	✓		✓			✓	✓
Walls Description	✓	✓	✓	✓	✓	✓	✓
Floor Description	✓	✓	✓	✓	✓	✓	✓
Windows Energy Efficiency	✓	✓	✓				
Windows Environmental Efficiency		✓	✓				
Walls Energy Efficiency		✓	✓	✓	✓	✓	✓
Walls Environmental Efficiency	✓	✓	✓	✓		✓	✓
Roof Description	✓	✓	✓	✓	✓	✓	✓
Roof Energy Efficiency	✓	✓	✓		✓		
Roof Environmental Efficiency	✓	✓	✓				
Lighting Environmental Efficiency	✓						
Number of Heated Rooms				✓	✓	✓	✓
Number of Habitable Rooms				✓	✓	✓	✓
Wind speed [km/h] (Annual Average)				✓			
Pressure [Hg] (Annual Average)					✓	✓	

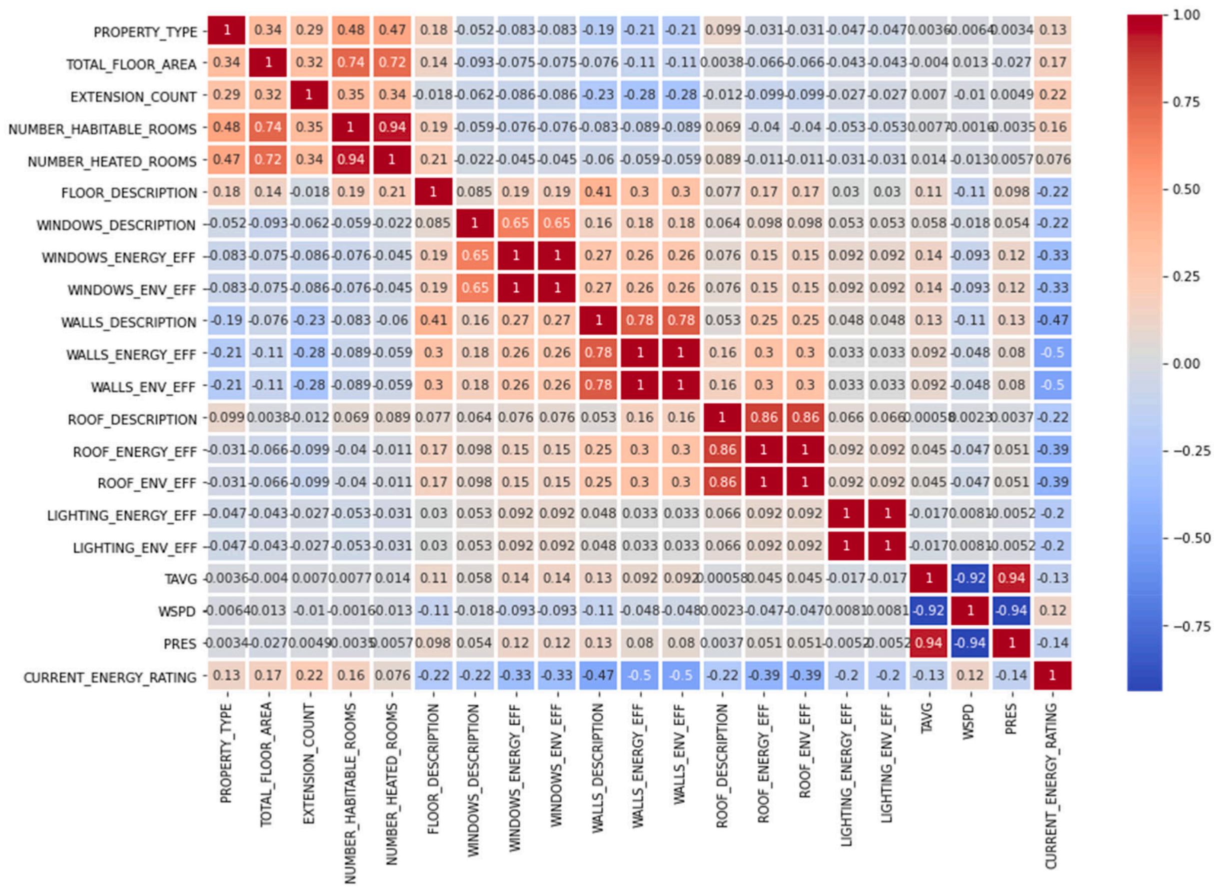

Fig. 6. Correlation between variables.

with themselves. The correlation values close to one represents a strong positive correlation between two features. Therefore, number of heated rooms has a strong correlation to number of habitable rooms. Windows environmental efficiency has a strong correlation to window energy efficiency. Also, Wall energy efficiency has a strong correlation to wall environmental efficiency.

3.5. Model development

Classification is a supervised learning method of developing a model to forecast the class label for an unseen instance. If the quantity of class labels connected with each instance is one, it is known as Single Label Classification (SLC). SLC is grouped into binary classification and multi-class classification. Binary classification is a method of classification that includes only two different class labels such as spam detection model (spam or not spam), result generation model (pass or fail). Subsequently, multi-class classification is a method of classification that includes more than two different class labels such as marital status model (married, single, divorced, widow/widower), and ethnicity detection (African American, European American, White British). This study employs the multi-class classification using the UK standard rating scale of A to G ('A' denoted as the most efficient and 'G' as the least efficient) as the class labels utilized in this study.

3.5.1. Hardware specification

The pre-processing of data, model training and testing were executed using python programming language. These computations were performed on an Apple M1 chip MacBook Air (OS = Big Sur Version = 11.4 and with RAM = 16 GB and 8 cores). Python libraries and packages (Pandas, numpy and scikit-learn) were utilized for the main core of this experiment.

3.5.2. Machine learning algorithms

Support Vector Machines: This is a machine learning method also identified as Support Vector Classifier (SVC), it is recognised as one of the most accurate techniques amongst data mining algorithm (Wu et al., 2008). SVM has gained more attention, owing to its capability of effectively generating good solutions to non-linear problems in diverse sample sizes (Chen et al., 2022; Hai-xiang Zhao & Magoulès, 2012b). SVM is based on the kernel, a method primarily computed for solving binary classification problems proposed by Vapnik in the early 1990s (Sonkamble & Doye, 2008). The parameters used in the model creation for the SVC model are: ('C' = 1.0; Epsilon = 0.1; Kernel = radial basis).

K Nearest Neighbors (KNN)**:** KNN algorithm is a non-parametric ML method that uses similarity or distance function to estimate results based on the k closest training samples in the feature space (Ortiz-Bejar et al., 2018). KNN is one of the most utilized distance functions methods that performs effectually on numerical data (Ali et al., 2019; Olu-Ajayi, 2017). The parameters utilized in the development of the KNN are 5 neighbors, 30 leaf and size uniform weights.

Random Forest: This is an ensemble technique developed based on the ensemble learning theory, which makes the learning of both simple and complicated problems achievable (Ahmad et al., 2017b); RF algorithm often generates good performance using default parameters. Hence, RF algorithm is obtaining increased recognition in the field of energy use prediction (Ahmad et al., 2017b; Chen et al., 2019; Fan et al., 2017; Z. Wang et al., 2018). The Random Forest (RF) model was developed with 10 estimators.

Decision Tree: This utilizes a tree-like flow chart to divide data into sets. It is a versatile method that can progress with an increased sample size (Domingos, 2012). In comparison with other ML techniques, DT is uncomplicated and easy to understand. Furthermore, the implementation of DT does not demand convoluted computational knowledge.

Though, the results often have clear deviations in their predictions from real outcomes (Yu et al., 2010). The Decision Tree (DT) was developed using the 'best' splitter parameter.

Gradient Boosting: This is a type of boosting method that develops models in phases, but it generalizes these by applying a random differentiable loss function (Flores & Keith, 2019). GB is one of the ML techniques that can be used for both classification and regression problems. The GB model was developed using certain parameters (Learning rate = 0.1, Loss function = deviance, estimators = 100).

Extra trees: This is a type of tree-based ensemble method that utilizes a decision tree as the key component with a top-down method. The extra trees algorithm is considered suitable when the utilized dataset contains a significant number of continuous variables, as it decreases the computational burden and arbitrarily chooses the best feature to split (Ravi, 2020). The ET model was developed using 100 estimators.

Multi-Layer Perceptron: This is a type of neural network that employs a feed forward propagation procedure with a single hidden layer where latent and abstract features are learned (Adegoke, 2019; Donoghue and Roantree, 2015). Neural networks are recognised for their good performance with large datasets as it requires adequate data to train the model (Bourhnane et al., 2020). The MLP model was developed using certain parameters (learning rate = 0.001, Activation = 'RELU' and solver = Adam).

Adaboost: This is a simple boosting method that is often utilized for resolving classification problems (Rahul et al., 2018). It is recognized for its high predictive speed and low time consumption (Aadithyan et al., 2020). The Ada boost model was developed using 1.0 learning rate.

Deep Neural Network: Deep neural network or deep learning is a machine learning technique that adopts deep patterns of the neural networks using multiple hidden layers (Olu-Ajayi, Alaka, Owolabi, Akanbi, & Ganiyu, 2023; Hoang and Kang, 2019). The regular neural networks consist of two to three layers, limiting their capabilities to expressing intricate functions (Lei et al., 2021). However, the deep neural network often consists of five or more layers of neural networks which enable the generation of better performance and increased accuracy. The developed deep learning model consists of five layers (1 input layer, 3 hidden layers and one output layer).

Bagging: Bagging is an abbreviation from Bootstrap Aggregating, which is one of the most utilized and famous amongst ensembles learning methods (Zeng et al., 2010). Bagging generates parallel various classifiers and then ensemble them, so it chooses specific base classifier algorithms to train base classifiers on random redistribution training datasets. The number of estimators were assigned to 10 in the development of the Bagging model.

Gaussian Naïve Bayes: In the data mining and machine learning field, this ML method is considered one of the valuable classification methods due to its effectiveness. However, they fail in the assumption of conditional independence among the features (Jahromi & Taheri, 2017). The default parameters were utilized in the development of the GNB.

Bernoulli Naïve Bayes: This is a type of multivariate method often utilized in classification tasks using binary independent features. It is popular in document classification tasks i.e., to detect if a term is under consideration or not (Singh et al., 2019). The BNB model was developed using 1.0 alpha.

Dummys: The dummy classifiers are methods that randomly guess the prediction classes, which can similarly attain a certain level of prediction accuracy (C. Wang et al., 2018).

Quadratic Discriminant Analysis: This is a common method for supervised classification, which gaussian distribution in modelling the likelihood of each class, then utilizes posterior distribution to predict the class (Srivastava et al., 2007). The default parameters were utilized in the development of the Quadratic Discriminant Analysis.

3.6. Model evaluation

Accuracy: This is a type of performance measure which considered the most used in evaluating classification tasks. It is the calculation of the exact match of the estimated values and real values. Also, this measure is often used considered to justify that the model developed is appropriate (Gonzalez-Abril et al., 2014). The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced Accuracy: Accuracy is known as the percentage of occurrences estimated correctly, while balanced accuracy is the mean of the accuracies for each class (Miller et al., 2012). There are two types of evaluation measures used to measure the balanced accuracy of the prediction result, namely sensitivity and specificity. The formula for balance accuracy is:

$$Sensitivity = \frac{TP}{TP + FN} Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{Sensitivity + Specificity}{2}$$

F1 score: This is a technique for computing the weighted average of the precision and recall, where the score close to one is considered the best and the score closest to zero is the worst. The formula for the F1 score is:

$$Precision = \frac{TP}{(TP + FP)} || Recall = \frac{TP}{(TP + FN)}$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

ROC AUC: This is the area under the Receiver Operator Characteristic (ROC) curve. It is largely acknowledged as one of the most appropriate pointers for the classification performance. The best ROC AUC score signifying outstanding accuracy is one. However, the lowest ROC AUC score is 0.5 (Egwim et al., 2021).

4. Result and discussion

This study conducts a reliability analysis to investigate the effect of sample size on classification performance. Five different percentages (i.e., 20%, 40%, 60%, 80% and 100%) of data were utilized as training and testing sets to develop energy use rating prediction models. Fourteen machine learning models were developed using five percentages of data. The most effective predictive model is determined in the five cases of data availability using the model performance measures as stated in section 3.5. The performance values of accuracy closest to one are considered the most effective model. As summarized in Table 3, the Gradient Boosting (GB) model produced the best predictions in five cases with accuracy achieved about 0.66–0.68 compared with Random Forest (RF) which also emanated the second best in five cases 0.64–0.66, 0.62–0.63 with Extra trees and 0.60–0.62% for K-Nearest Neighbour (KNN). The least effective predictive models are Quadratic Discriminant Analysis (QDA), Gaussian Naïve Bayes (GNB) and Adaboost. Fig. 7 shows the prediction performance distribution of each model using different data sizes.

The change in data size had effects on the performance of the different algorithms except for the dummy and bagging classifier, which remained constant in the five cases. It is noted that Gradient Boosting (GB) outperformed other models using 20% data and there appears no significant changes in the prediction performance (accuracy) with the increased sample size. This suggests that the size of the data has no direct impact on the predictive accuracy. Therefore, 20% and larger data can be considered sufficient for energy use prediction using Gradient Boosting (GB). However, further investigation explores the relative

Table 3
Performance result for each model using five data sizes.

Data Size	20% Data					40% Data					60% Data				
Model	**Training Time**	**ROC AUC**	**Accuracy**	**Balanced Accuracy**	**F1**	**Training Time**	**ROC AUC**	**Accuracy**	**Balanced Accuracy**	**F1**	**Training Time**	**ROC AUC**	**Accuracy**	**Balanced Accuracy**	**F1**
Support Vector Machines	19.32	0.79	0.46	0.14	0.29	75.06	0.78	0.46	0.14	0.29	166.29	0.80	0.47	0.14	0.30
K Nearest Neighbors	0.63	0.79	0.62	0.38	0.61	2.20	0.78	0.61	0.34	0.60	4.31	0.80	0.60	0.35	0.59
Random Forest	0.64	0.85	0.65	0.47	0.64	1.29	0.84	0.66	0.43	0.65	2.06	0.84	0.65	0.44	**0.64**
Decision Tree	0.05	0.70	0.60	0.48	0.60	0.13	0.70	0.59	**0.48**	0.60	0.21	0.67	0.58	0.42	0.58
Gradient Boosting	5.36	**0.90**	**0.68**	**0.50**	**0.67**	12.98	**0.87**	**0.67**	0.44	**0.66**	21.16	**0.85**	**0.66**	0.38	0.65
Extra trees	0.68	0.84	0.63	0.47	0.63	1.38	0.81	0.63	0.41	0.62	2.13	0.80	0.62	0.40	0.61
Multi-Layer Perceptron	0.80	0.66	0.44	0.25	0.31	1.02	0.71	0.43	0.25	0.29	1.31	0.79	0.52	0.26	0.49
Adaboost	1.12	0.62	0.46	0.22	0.38	2.30	0.64	0.46	0.27	0.37	3.16	0.63	0.41	0.32	0.43
Deep Neural Network	2.34	0.82	0.48	0.18	0.34	6.93	0.83	0.59	0.22	0.54	15.92	**0.85**	0.58	0.24	0.55
Bagging	30.31	0.50	0.46	0.14	0.29	113.58	0.50	0.46	0.14	0.29	265.09	0.50	0.47	0.14	0.30
Gaussian Naïve Bayes	0.03	0.76	0.39	0.39	0.33	0.05	0.74	0.38	0.38	0.45	0.07	0.80	0.44	0.42	0.48
Bernoulli Naïve Bayes	0.03	0.81	0.54	0.30	0.52	0.05	0.80	0.54	0.30	0.53	0.08	0.80	0.54	0.30	0.53
Dummy	0.00	0.50	0.46	0.14	0.29	0.00	0.50	0.46	0.14	0.29	0.00	0.50	0.47	0.14	0.30
Quadratic Discriminant Analysis	0.05	0.70	0.31	0.26	0.32	0.09	0.74	0.48	0.30	0.46	0.13	0.70	0.14	0.27	0.18

Data Size	80.00					100.00				
Model	**Training Time**	**ROC AUC**	**Accuracy**	**Balanced Accuracy**	**F1**	**Training Time**	**ROC AUC**	**Accuracy**	**Balanced Accuracy**	**F1**
Support Vector Machines	294.11	0.78	0.46	0.14	0.29	496.46	0.78	0.46	0.14	0.29
K Nearest Neighbors	7.76	0.78	0.60	0.34	0.58	12.70	0.78	0.61	0.31	0.59
Random Forest	2.90	0.83	0.64	0.41	0.63	3.96	0.82	0.64	0.37	0.63
Decision Tree	0.30	0.68	0.58	0.45	0.58	0.41	0.65	0.58	0.39	0.58
Gradient Boosting	30.42	**0.86**	**0.67**	0.40	**0.65**	40.93	**0.86**	**0.67**	0.38	**0.65**
Extra trees	2.96	0.80	0.62	0.39	0.61	3.88	0.79	0.62	0.34	0.61
Multi-Layer Perceptron	5.06	0.78	0.39	0.24	0.31	5.81	0.78	0.56	0.19	0.48
Adaboost	4.20	0.63	0.44	0.32	0.45	5.35	0.66	0.47	0.34	0.47
Deep Neural Network	14.98	0.84	0.60	0.27	0.59	33.89	**0.86**	0.63	0.27	0.60
Bagging	471.34	0.50	0.46	0.14	0.29	755.18	0.50	0.46	0.14	0.29
Gaussian Naïve Bayes	0.10	0.68	0.13	0.35	0.16	0.13	0.73	0.16	0.36	0.20
Bernoulli Naïve Bayes	0.10	0.79	0.50	0.29	0.46	0.13	0.78	0.49	0.28	0.44
Dummy	0.00	0.50	0.46	0.14	0.29	0.00	0.50	0.46	0.14	0.29
Quadratic Discriminant Analysis	0.18	0.81	0.51	0.27	0.46	0.23	0.83	0.55	0.29	0.50

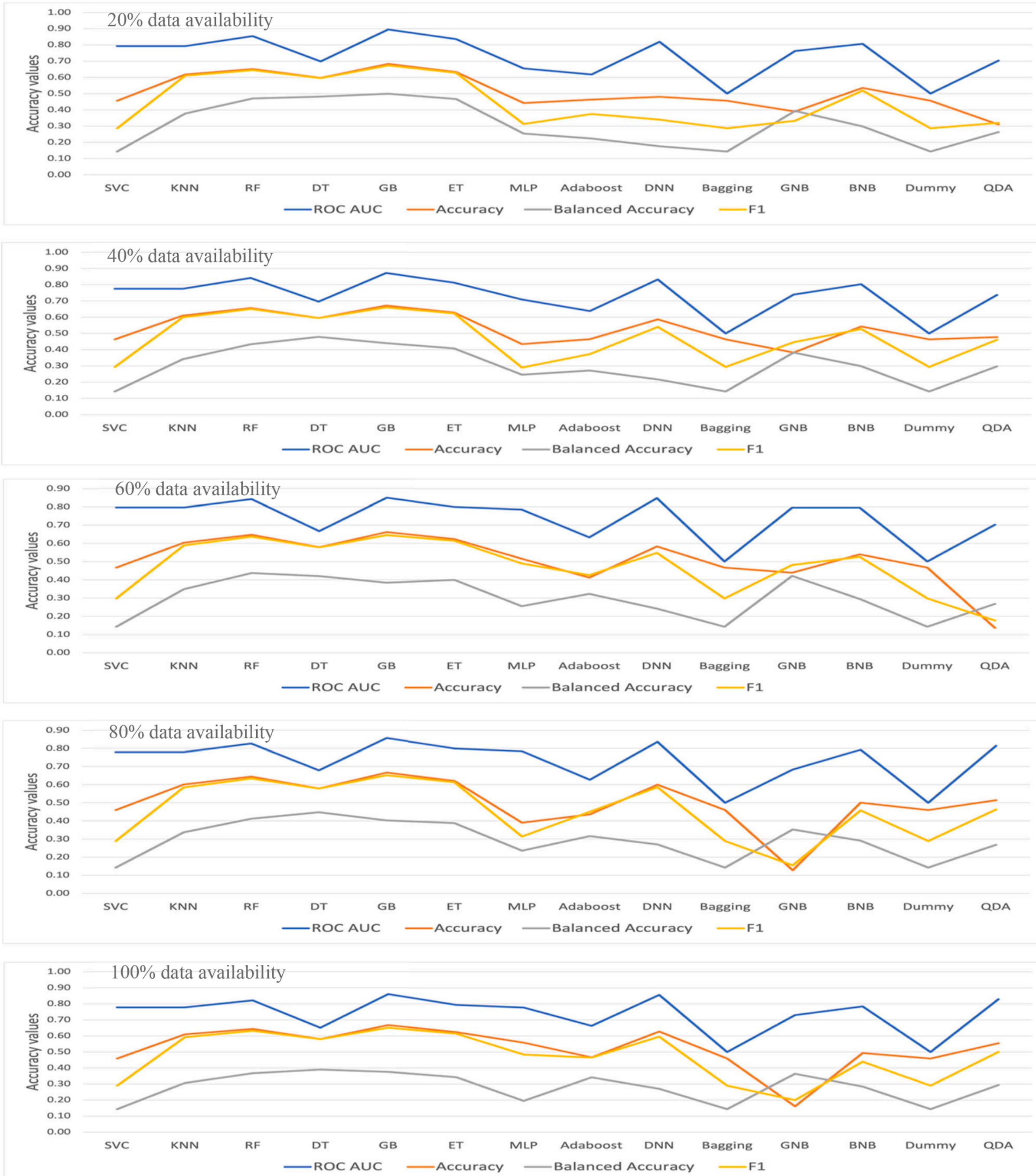

Fig. 7. Prediction performance distribution using different data sizes for each ML algorithm.

effect of the input features on the predictive performance(s).

The fourteen ML model's performance distribution is shown in Fig. 7 using a stacked line chart. This chart was employed to display a more comprehensible means to recognise the variation in model performance using different data sizes. Most of the ML models (such as SVC, KNN, RF, DT, among others) shows no significant difference in the different sizes of data used, as shown in Fig. 7. However, Deep Neural Network (DNN) shows a relative increase in performance based on increased data size. On the other hand, Gaussian Naïve Bayes (GNB) shows a decline in prediction performance based on the increase in data. Gradient boosting is the most effective model for energy use prediction, but there still lies no significant difference using different data sizes, with an achieved prediction accuracy of 0.66–0.68. Although DNN is not recognised as the most effective model, it does show an increase in performance based on data increase. This does suggest that the positive or negative effect of data availability or large data is predicated on the type of algorithm used.

To examine the impact of the input features on the predictive outcome. Seven feature selection methods were implemented on 100% data availability. The ten most relevant input features based on the feature selection methods were used to develop fourteen machine learning models. The result was further investigated to identify the most effective feature selection method and pinpoint the paramount features for energy use prediction. Figs. 8a to 8c displays the model's prediction performance based on input features selected using the different feature selection methods. Gradient Boosting (GB) produced the best result based on accuracy using Random Forest (RF), Permutation Importance, RFE and Extratree.

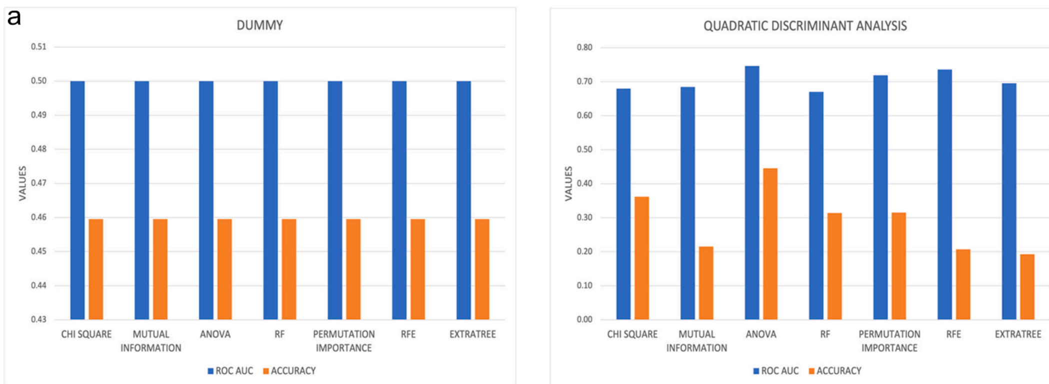

Fig. 8a. Models performance results using various FS methods.

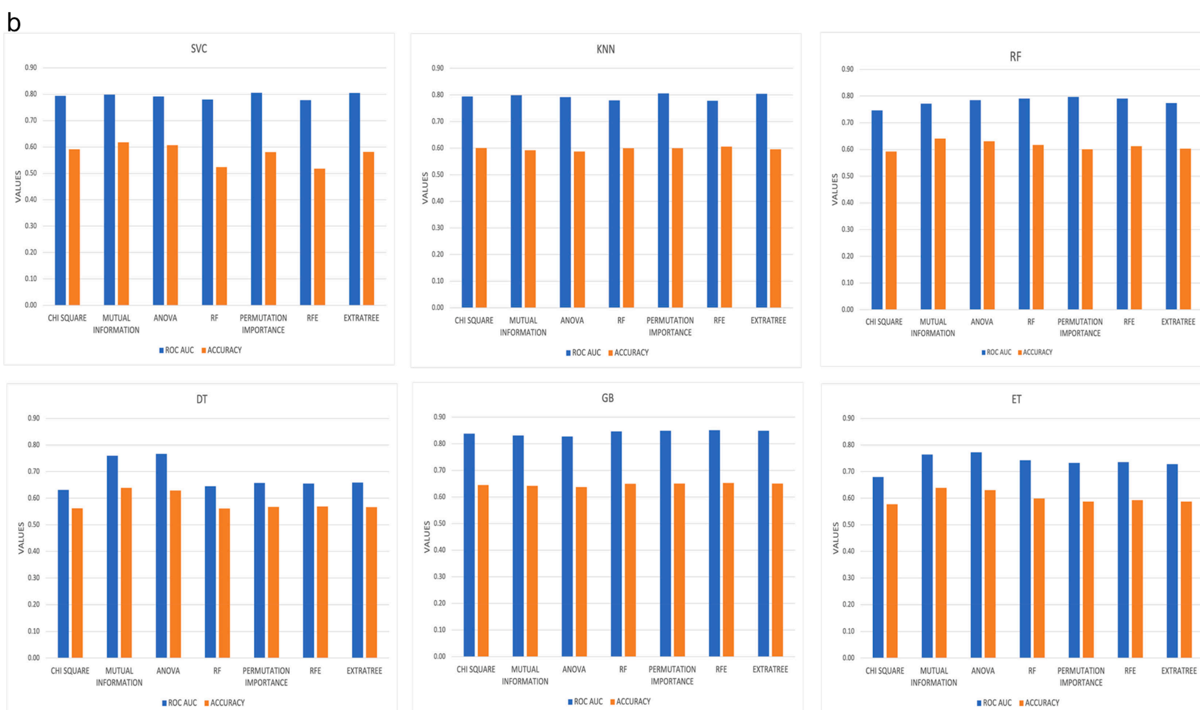

Fig. 8b. Models performance results using various FS methods.

Each model produced its highest accuracy using specific feature selection methods. For example, SVC produced good accuracy of 0.61–0.62 using Mutual Information and ANOVA. Likewise, RF genrated good predictive accuracy of 0.63–0.64 using Mutual Information and ANOVA. Similarly, DT and Extratree engender a good level of accuracy using Mutual Information and ANOVA. Furthermore, GaussianNB generated good predictive accuracy of 0.54 using RFE and 0.55 of BernoulliNB using RF and Extratree. On the other hand, all feature selection methods have no effect on the Dummy model, as it generated the same level of accuracy across all seven feature selection methods, as shown in Fig. 8a below.

Given that feature selection could have unfavourable impacts on certain algorithms (Balogun et al., 2021; Olu-Ajayi et al., 2022b), there is a need to compare each model with and without FS methods. Fig. 9 demonstrates a clear comparison of the negative and positive impacts of the FS techniques on the predictive accuracy of the fourteen ML models. Table 4 summarizes the prediction accuracy of each model with and without FS. Although the Gradient Boosting (GB) model emerged the best with and without feature selection, it achieved a better accuracy of 0.67 without feature selection as compared to 0.65 with feature selection. Also, Quadratic Discriminant Analysis (QDA) generated better predictive accuracy of 0.55 without FS and 0.45 with FS.

Conversely, Feature Selection had favourable impacts on certain ML algorithms. For example, DT achieved better accuracy 0.64 using feature selection and 0.58 without feature selection. MLP generated 0.64 predictive accuracy with FS and 0.56 without FS. This also applies to other models such as DNN, Bagging, GaussianNB among others. This level of increase and decrease based on feature selection is displayed in Fig. 9 below.

5. Implications of research

This study presents the theoretical and practical implications of this research.

5.1. Theoretical Implications

This study explored the utilization of various FS techniques in developing several machine learning models to conduct a clear and unbiased comparison. Some studies suggest that feature selection is

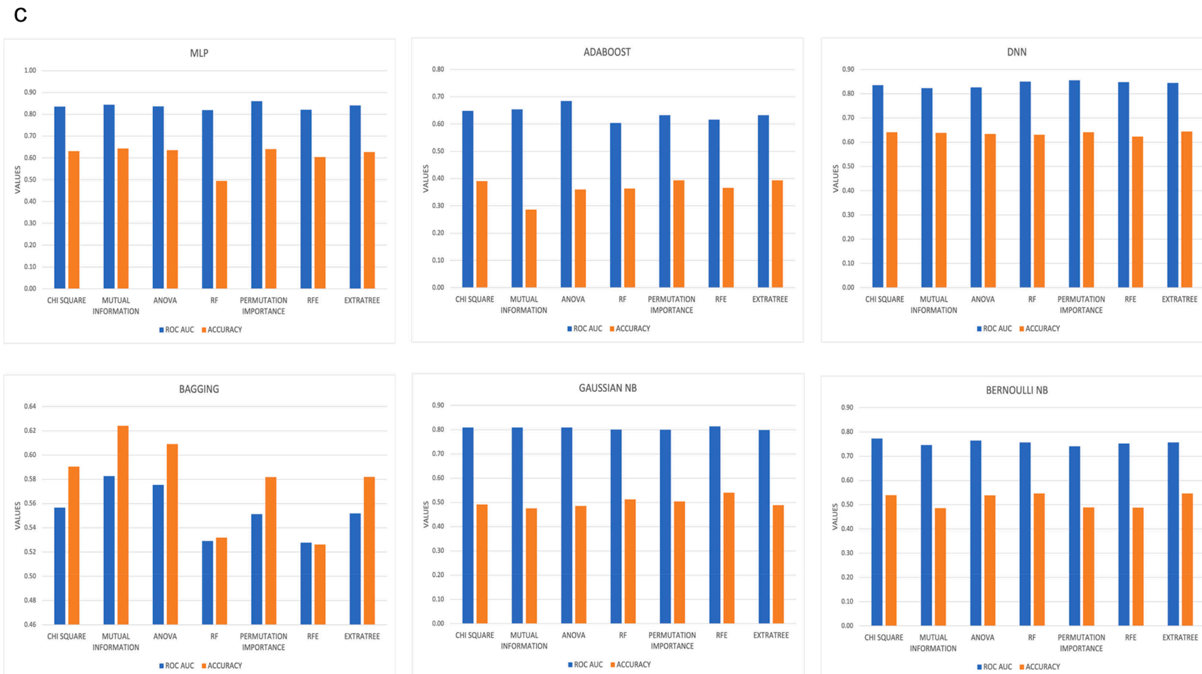

Fig. 8c. Models performance results using various FS methods.

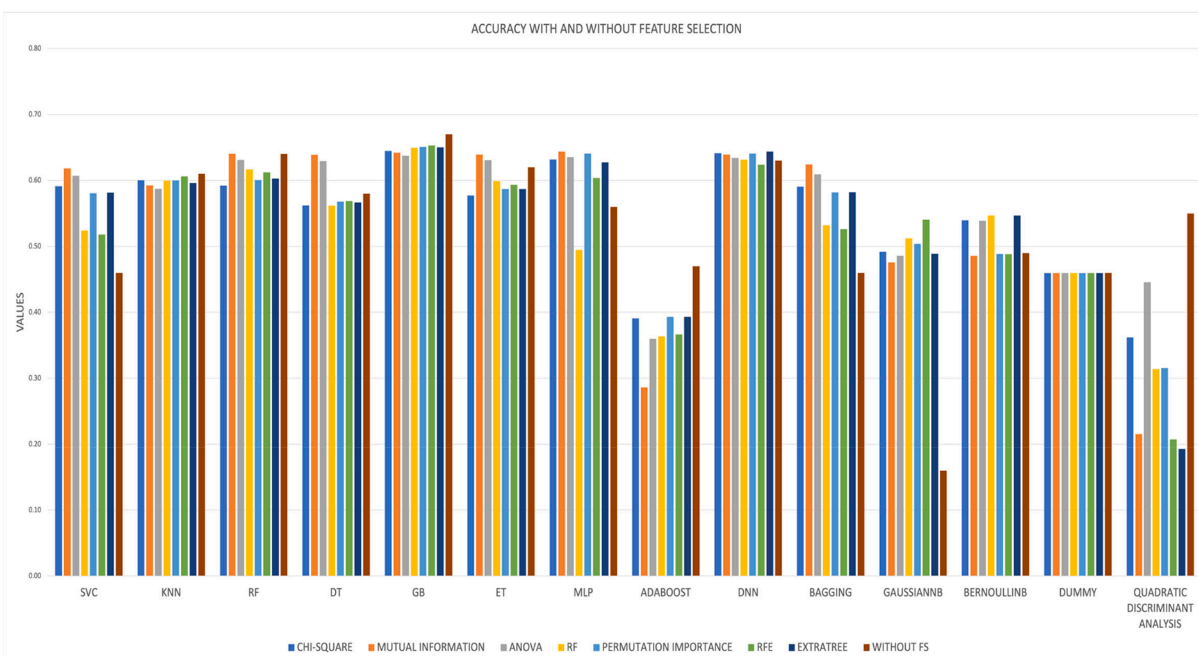

Fig. 9. Prediction performance distribution using different data sizes for each ML algorithm.

essential for optimum performance of the model (Alaka et al., 2018; Balogun et al., 2021; Zhang & Wen, 2019a; Hai-xiang Zhao & Magoulès, 2012a); some studies suggest that feature selection is unfavourable to certain algorithms (Alaka et al., 2019, 2018; Olu-Ajayi et al., 2022b). However, this study shows that feature selection can have a positive or negative impact on the predictive accuracy, depending on the algorithm selected. This supports previous research, which states that the achievement of a good predictive accuracy of a model is highly predicated on the type of algorithm and feature selection method chosen (Olu-Ajayi et al., 2022a). Therefore, if an appropriate feature selection method is not selected for the specific ML algorithm used, feature selection will not result in good accuracy. For example, SVC produced good results for Mutual Information and ANOVA, which are both filter methods. This could suggest that SVC is best suited with filter methods. However, some models achieved better accuracy without feature selection such as Gradient Boosting (GB) and Adaboost among others.

One general hypothesis in the machine learning world, states that the larger the data utilized for model training, the better the performance (Dalal, 2018; Goyal et al., 2020; Kaur & Gupta, 2017; Lee et al., 2011). In this study, the reliability analysis was conducted to substantiate or disprove this hypothesis. Different machine learning models were developed using five sample sizes (20%, 40%, 60%, 80%, 100%). This

Table 4
Performance result for each model with and without feature selection methods.

	Chi-Square	Mutual Information	ANOVA	RF	Permutation Importance	RFE	Extratree	Without FS
SVC	0.59	0.62	0.61	0.52	0.58	0.52	0.58	0.46
KNN	0.60	0.59	0.59	0.60	0.60	0.61	0.60	0.61
RF	0.59	0.64	0.63	0.62	0.60	0.61	0.60	0.64
DT	0.56	0.64	0.63	0.56	0.57	0.57	0.57	0.58
GB	0.64	0.64	0.64	0.65	0.65	0.65	0.65	0.67
ET	0.58	0.64	0.63	0.60	0.59	0.59	0.59	0.62
MLP	0.63	0.64	0.64	0.49	0.64	0.60	0.63	0.56
Adaboost	0.39	0.29	0.36	0.36	0.39	0.37	0.39	0.47
DNN	0.64	0.64	0.63	0.63	0.64	0.62	0.64	0.63
Bagging	0.59	0.62	0.61	0.53	0.58	0.53	0.58	0.46
GaussianNB	0.49	0.48	0.49	0.51	0.50	0.54	0.49	0.16
BernoullinB	0.54	0.49	0.54	0.55	0.49	0.49	0.55	0.49
Dummy	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
Quadratic Discriminant Analysis	0.36	0.22	0.45	0.31	0.32	0.21	0.19	0.55

finding engendered from the analysis suggests that the larger the data does not always lead to a better result. Although it is notable that a single study is not enough to substantiate this conclusion, and this should be subject to further investigation. Some studies have corroborated certain conclusions, for example, Deep Neural Networks (DNN) showed a relative increase in predictive accuracy based on the increase in data size. Past studies have shown that large data has an effect on certain machine learning algorithms such as neural networks, (Amasyali & El-Gohary, 2018; Bourhnane et al., 2020). Also, the study by Bourhnane et al., 2020 stipulates that neural network algorithms are dominant with big datasets, as they require sufficient data to train the model. On the other hand, SVC shows no significant increase in predictive accuracy based on the increase in the data size. This is can be subject to the conclusion that Support Vector Machines (SVM) are recognized for its ability to deliver good results effectively regardless of data size (Li et al., 2009; Qiong Li et al., 2010). Therefore, the general hypothesis on large data and better predictive performance is only peculiar to certain machine learning algorithms.

5.2. Practical Implications

The identification of the most relevant features that influence the energy performance of a building is important for several reasons and at various stages. During the development of a building energy prediction model, the utilization of only the relevant variables can improve the accuracy of the model. This can also help in reducing the model's complexity and avoiding overfitting. Additionally, this could reduce the time-consuming effort and high cost required to collect data on all variables that may affect building energy consumption. Furthermore, the use of only the relevant variables for an energy prediction model can reduce the computational cost. Moreover, identifying of the most relevant features is also imperative at the design stage of a building because this enables the building designer to discern which building features require optimization to achieve a low potential energy consumption outcome. Overall, the choice of the most suitable feature selection method can help in identifying the most relevant features that contribute to the energy use of a building. This study identifies the most suitable FS method for specific ML algorithms, for instance, SVC, ET, MLP and bagging were found to be most suited with filter feature selection methods such as chi-square, mutual information, and ANOVA, among others. In terms of technical, social and economic implications, technical expertise is required to implement different feature selection methods and exploring various FS methods to identify the most suitable for an ML algorithm it is more labour-intensive. This study delivers the most suitable FS method for certain ML algorithms, reducing the time-consuming process of exploring various FS methods. In real-world situations, the process of collecting and utilizing data could raise privacy concerns, particularly with personal data such as occupancy details among others. The data of numerous features are often collected However, in most cases, only a few of these features may be related to the target output (Kira and Rendell, 1992).The identification of the most relevant feature will help limit the types of data required and reduce the cost of data collection. Developing a high performing building energy consumption prediction model can help organizations optimize the use of energy resources, leading to cost savings and improved sustainability. Additionally, the implementation of feature selection and reliability analysis in this study, can help model developers identify the ML algorithms that are sensitive to changes in the data or feature size.

6. Conclusion and recommendation

Contrary to the popular theory that feature selection is required to achieve better accuracy (Alaka et al., 2018; Balogun et al., 2021; Olu-Ajayi et al., 2021; Zhang & Wen, 2019a; Hai-xiang Zhao & Magoulès, 2012a), this study further investigated the utilization of seven feature selection methods on fourteen machine learning algorithms in predicting energy use in buildings. Although it is noted that in most cases, FS does improve the performance of a model, results show that feature selection can have a positive or negative impact on the model's performance. Therefore, it is concluded that the achievement of a good predictive accuracy of a model is predicated on the choice of an appropriate feature selection method. However, in this study, it is noted that some models perform better without feature selection such as Gradient Boosting (GB), Quadratic Discriminant Analysis (QDA) among others. Gradient Boosting (GB) produced the highest accuracy for building energy performance prediction without feature selection and using different sizes of data.

The implementation of a reliability analysis was conducted to satisfy the second objective. It is concluded that the larger the data does not necessarily lead to more accurate results as opposed to previous studies (Dalal, 2018; Goyal et al., 2020; Kabir, 2020). However, this hypothesis is true in the utilization of certain ML algorithms such as DNN. Therefore, the utilization of a larger data size to train a model can lead to better accuracy dependent on the algorithm selected. Thus, this hypothesis is considered true, however it is not generalizable. While it has been deduced that there is no specific FS technique considered faultless and favourable to all classification algorithms in the study, the filter feature selection methods are considered the most suitable method in developing ML classification models, this is because the filter feature selection methods achieved the best accuracy in 9 out of 14 energy prediction models.

Considering various FS techniques have led to good model performance at various times (Ahmad et al., 2017a; Dong et al., 2021; Zhang & Wen, 2019b) and it is widely acknowledged that one study is not adequate to justify the aforementioned conclusion. Therefore, future study should investigate different other FS methods in comparison with the filter FS methods for predicting building energy performance. Furthermore, future studies should also explore other machine learning classification algorithms on different sizes of an even larger dataset.

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.

References

Aadithyan, V., Goud, T. S. S., Reddy, G. K., Chaitanya, P. N., Surya, V. J., & Rao, K. P. (2020). *Smart Face Recognition System.*

Adegoke, M., Hafiz, A., Ajayi, S., & Olu-Ajayi, R. (2022). Application of Multilayer Extreme Learning Machine for Efficient Building Energy Prediction. *Energies, 15*, 9512.

Adegoke, Muideen, et al. (2019). Two noise tolerant incremental learning algorithms for single layer feed-forward neural networks. *Journal of Ambient Intelligence and Humanized Computing*, 1–15.

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., & Ajayi, S. (2019). A big data analytics approach for construction firms failure prediction models. *IEEE Transactions on Engineering Management, 66*, 689–698. https://doi.org/10.1109/TEM.2018.2856376

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications, 94*, 164–184. https://doi.org/10.1016/j.eswa.2017.10.040

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Science, 1*, 1559. https://doi.org/10.1007/s42452-019-1356-9

Allouhi, A., El Fouih, Y., Kousksou, T., Jamil, A., Zeraouli, Y., & Mourad, Y. (2015). Energy consumption and efficiency in buildings: Current status and future trends. *Journal of Cleaner Production, Special Issue: Toward a Regenerative Sustainability Paradigm for the Built Environment: From vision to reality, 109*, 118–130. https://doi.org/10.1016/j.jclepro.2015.05.139

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews, 81*, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Amiri, F., Rezaei Yousefi, M., Lucas, C., Shakery, A., & Yazdani, N. (2011). Mutual information-based feature selection for intrusion detection systems. *Journal of Network and Computer Applications, Advanced Topics in Cloud Computing, 34*, 1184–1199. https://doi.org/10.1016/j.jnca.2011.01.002

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*, 971–989. https://doi.org/10.1109/TCBB.2015.2478454

Asir, D., Gnana, A., & Leavline, E. J. (2016). *Literature Review on Feature Selection Methods for High-Dimensional Data.*

Aziz, R., Verma, C. K., Srivastava, N., Aziz, R., Verma, C. K., & Srivastava, N. (2017). Dimension reduction methods for microarray data: A review. *AIMSBOA, 4*, 179–197. https://doi.org/10.3934/bioeng.2017.1.179

Bagnasco, A., Fresi, F., Saviozzi, M., Silvestro, F., & Vinci, A. (2015). Electrical consumption forecasting in hospital facilities: An application case. *Energy and Buildings, 103*, 261–270. https://doi.org/10.1016/j.enbuild.2015.05.056

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences, 32*, 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Balogun, H., Alaka, H., & Egwim, C. N. (2021). Boruta-grid-search least square support vector machine for NO2 pollution prediction using big data analytics and IoT emission sensors. *Applied Computing and Informatics ahead-of-print.* https://doi.org/10.1108/ACI-04-2021-0092

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks, 5*, 537–550. https://doi.org/10.1109/72.298224

BEIS [WWW Document]. (2019). GOV.UK. URL https://www.gov.uk/government/publications/beis-annual-report-and-accounts-2019-to-2020 (accessed 9.15.21).

Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine, 195*, 105635. https://doi.org/10.1016/j.cmpb.2020.105635

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, Relevance, 97*, 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis, 143*, 106839. https://doi.org/10.1016/j.csda.2019.106839

Bourhnane, S., Abid, M. R., Lghoul, R., Zine-Dine, K., Elkamoun, N., & Benhaddou, D. (2020). Machine learning for energy consumption prediction and scheduling in smart buildings. *SN Applied Science, 2*, 297. https://doi.org/10.1007/s42452-020-2024-9

Survey, B. E. E. (2016). *Building Energy Efficiency Survey (BEES) [WWW Document]*. GOV: UK. https://www.gov.uk/government/collections/non-domestic-buildings-energy-use-project (accessed 3.29.21.

Canales, V. R. (2016). *Using a Supervised Learning Model: Two-Class Boosted Decision Tree Algorithm for Income Prediction 7.*

Carrera, B., Peyrard, S., & Kim, K. (2021). Meta-regression framework for energy consumption prediction in a smart city: A case study of Songdo in South Korea. *Sustainable Cities and Society, 72*, 103025. https://doi.org/10.1016/j.scs.2021.103025

Castelli, M., Trujillo, L., Vanneschi, L., & Popovič, A. (2015). Prediction of energy performance of residential buildings: A genetic programming approach. *Energy and Buildings, 102*, 67–74. https://doi.org/10.1016/j.enbuild.2015.05.013

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40th-year commemorative issue, 40*, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, Y., Guo, M., Chen, Z., Chen, Z., & Ji, Y. (2022). Physical energy and data-driven models in building energy prediction: A review. *Energy Reports, 8*, 2656–2671. https://doi.org/10.1016/j.egyr.2022.01.162

Chen, Y.-T., Piedad, E., & Kuo, C.-C. (2019). Energy consumption load forecasting using a level-based random forest classifier. *Symmetry, 11*, 956. https://doi.org/10.3390/sym11080956

Curtis, J., Devitt, N., & Whelan, A. (2014). Estimating Building Energy Ratings for the Residential Building Stock: Location and Occupancy. *Papers, Papers, No. WP489*.

Dalal, K. R. (2018). Review on Application of Machine learning Algorithm for Data Science. In: *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*. IEEE, pp. 270–273.

Dandotiya, B. (2020). *Climate-Change-and-Its-Impact-on-Terrestrial-Ecosystems.* https://doi.org/10.4018/978-1-7998-3343-7.ch007.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*, 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5

Ding, H., Feng, P.-M., Chen, W., & Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems, 10*, 2229–2235. https://doi.org/10.1039/C4MB00316K

Ding, Y., & Liu, X. (2020). A comparative analysis of data-driven methods in building energy benchmarking. *Energy and Buildings, 209*, 109711. https://doi.org/10.1016/j.enbuild.2019.109711

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*, 78–87. https://doi.org/10.1145/2347736.2347755

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. https://doi.org/10.1016/j.enbuild.2004.09.009

Dong, Z., Liu, J., Liu, B., Li, K., & Li, X. (2021). Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. *Energy and Buildings, 241*, 110929. https://doi.org/10.1016/j.enbuild.2021.110929

Donoghue, J. O., & Roantree, M. (2015). A Framework for Selecting Deep Learning Hyper-parameters. In: Maneth, S. (Ed.), *Data Science, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 120–132.

Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics, 61*, 101224. https://doi.org/10.1016/j.ecoinf.2021.101224

Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., & Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications, 6*, 100166. https://doi.org/10.1016/j.mlwa.2021.100166

Faisal, H. M., Javaid, N., Sarfraz, B., Baqi, A., Bilal, M., Haider, I., & Shuja, S. M. (2019). Prediction of Building Energy Consumption Using Enhance Convolutional Neural Network. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Eds.), *Web, Artificial Intelligence and Network Applications, Advances in Intelligent Systems and Computing* (pp. 1157–1168). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-15035-8_111.

Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy, 127*, 1–10. https://doi.org/10.1016/j.apenergy.2014.04.016

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy, 195*, 222–233. https://doi.org/10.1016/j.apenergy.2017.03.064

Fathi, S., Srinivasan, R., Fenner, A., & Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews, 133*, 110287. https://doi.org/10.1016/j.rser.2020.110287

Feng, C., & Zhang, J. (2020). Assessment of aggregation strategies for machine-learning based short-term load forecasting. *Electric Power Systems Research, 184*. https://doi.org/10.1016/j.epsr.2020.106304

Flores, V., & Keith, B. (2019). Gradient boosted trees predictive models for surface roughness in high-speed milling in the steel and aluminum metalworking industry. *Complexity, 2019*, e1536716.

Gonzalez-Abril, L., Nuñez, H., Angulo, C., & Velasco, F. (2014). GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems. *Applied Soft Computing, 17*, 23–31. https://doi.org/10.1016/j.asoc.2013.12.013

Goyal, K., Tiwari, N., & Sonekar, J. (2020). An anatomization of data classification based on machine learning techniques. *IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 7*, 713–716.

Hoang, D. T., & Kang, H. J. (2019). Rotary Machine Fault Diagnosis Using Scalogram Image and Convolutional Neural Network with Batch Normalization. In Huang, D.-S., Huang, Z.-K., Hussain, A. (Eds.), *Intelligent Computing Methodologies, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 283–293. https://doi.org/10.1007/978-3-030-26766-7_26.

Hsu, D. (2015). Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy, 83*, 144–155. https://doi.org/10.1016/j.energy.2015.02.008

Iqbal, M., Muneeb Abid, M., Noman, M., & Manzoor, Engr. Dr. A. (2020). Review of feature selection methods for text classification. *International Journal of Advanced Computer Research, 10*, 2277–7970. https://doi.org/10.19101/IJACR.2020.1048037

Jahromi, A. H., & Taheri, M. (2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In *2017 Artificial Intelligence and Signal Processing Conference (AISP). Presented at the 2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pp. 209–212. https://doi.org/10.1109/AISP.2017.8324083.

Jing, W., Zhen, M., Guan, H., Luo, W., & Liu, X. (2022). A prediction model for building energy consumption in a shopping mall based on Chaos theory. *Energy Reports, 8*, 5305–5312. https://doi.org/10.1016/j.egyr.2022.03.205

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Presented at the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458.

Kabir, M. A. (2020). Vehicle speed prediction based on road status using machine learning. *Advanced Research in Energy and Engineering, 2*.

Kapetanakis, D.-S., Mangina, E., & Finn, D. P. (2017). Input variable selection for thermal load predictive models of commercial buildings. *Energy and Buildings, 137*, 13–26. https://doi.org/10.1016/j.enbuild.2016.12.016

Kaur, K., & Gupta, O. P. (2017). A machine learning approach to determine maturity stages of tomatoes. *Oriental journal of computer science and technology, 10*, 683–690.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Sleeman, D., & Edwards, P. (Eds.), *Machine Learning Proceedings 1992*. Morgan Kaufmann, San Francisco (CA), pp. 249–256. https://doi.org/10.1016/B978-1-55860-247-2.50037-1.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, Relevance, 97*, 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International journal of computer science, 1*, 111–117.

Kunasekaran, K. K. H., & Sugumaran, R. (2016). Exploratory analysis of feature selection techniques in medical image processing. *Medical Image Processing, 5*.

Kusiak, A., Li, M., & Zhang, Z. (2010). A data-driven approach for steam load prediction in buildings. *Applied Energy, 87*, 925–933. https://doi.org/10.1016/j.apenergy.2009.09.004

Lee, S., Kim, Changmin, Park, Y., Son, H., & Kim, Changwan (2011). *Data Mining-Based Predictive Model to Determine Project Financial Success Using Project Definition Parameters*.

Lei, L., Chen, W., Wu, B., Chen, C., & Liu, W. (2021). A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings, 240*, 110886. https://doi.org/10.1016/j.enbuild.2021.110886

Li, C., Tao, Y., Ao, W., Yang, S., & Bai, Y. (2018). Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition. *Energy, 165*, 1220–1227. https://doi.org/10.1016/j.energy.2018.10.113

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys, 50, 94: 1 –94*, 45. https://doi.org/10.1145/3136625

Li, K., Xie, X., Xue, W., Dai, X., Chen, X., & Yang, X. (2018). A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. *Energy and Buildings, 174*, 323–334. https://doi.org/10.1016/j.enbuild.2018.06.017

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. https://doi.org/10.1016/j.enconman.2008.08.033

Liu, Y., Chen, H., Zhang, L., Wu, X., & Wang, X. (2020). Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China. *Journal of Cleaner Production, 272*, 122542. https://doi.org/10.1016/j.jclepro.2020.122542

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences, Special Section on High Order Fuzzy Sets, 179*, 2208–2217. https://doi.org/10.1016/j.ins.2009.02.014

Marino, C., Nucara, A., & Pietrafesa, M. (2017). Does window-to-wall ratio have a significant effect on the energy consumption of buildings? A parametric analysis in Italian climate conditions. *Journal of Building Engineering, 13*, 169–183. https://doi.org/10.1016/j.jobe.2017.08.001

Marwan, M. (2020). The effect of wall material on energy cost reduction in building. *Case Studies in Thermal Engineering, 17*, 100573. https://doi.org/10.1016/j.csite.2019.100573

Miller, Z., Dickinson, B., & Hu, W. (2012). *Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features*. https://doi.org/10.4236/ijis.2012.224019

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry, 132*, 116045. https://doi.org/10.1016/j.trac.2020.116045

Newgard, C. D., & Lewis, R. J. (2015). Missing data: How to best account for what is not known. *Journal of the American Medical Association, 314*, 940–941. https://doi.org/10.1001/jama.2015.10516

Olu-Ajayi, R. (2017). An Investigation into the Suitability of k-Nearest Neighbour (k-NN) for Software Effort Estimation. ijacsa 8. https://doi.org/10.14569/IJACSA.2017.080628.

Olu-Ajayi, R., & Alaka, H. (2021). Building energy consumption prediction using deep learning. *Environmental Design and Management Conference (EDMIC)*.

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Grishikashvili, K., Sunmola, F., Oseghale, R., & Ajayi, S. (2021). Ensemble learning for energy performance prediction of residential buildings. *Environmental Design and Management Conference (EDMIC)*.

Olu-Ajayi, R., Alaka, H., Owolabi, H., Akanbi, L., & Ganiyu, S. (2023). Data-Driven Tools for Building Energy Consumption Prediction: A Review. *Energies, 16*, 2574. https://doi.org/10.3390/en16062574

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022a). Machine learning for energy performance prediction at the design stage of buildings. *Energy for Sustainable Development, 66*, 12–25. https://doi.org/10.1016/j.esd.2021.11.002

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022b). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering, 45*, 103406. https://doi.org/10.1016/j.jobe.2021.103406

Ortiz-Bejar, José, Graff, M., Tellez, E. S., Ortiz-Bejar, Jesús, & Jacobo, J. C. (2018). k-Nearest Neighbor Regressors Optimized by using Random Search. In *2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). Presented at the 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, pp. 1–5. https://doi.org/10.1109/ROPEC.2018.8661399.

Paudel, S., Elmitri, M., Couturier, S., Nguyen, P. H., Kamphuis, R., Lacarrière, B., & Le Corre, O. (2017). A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. *Energy and Buildings, 138*, 240–256. https://doi.org/10.1016/j.enbuild.2016.11.009

Pham, A.-D., Ngo, N.-T., Ha Truong, T. T., Huynh, N.-T., & Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production, 260*, 121082. https://doi.org/10.1016/j.jclepro.2020.121082

Pirbazari, A. M., Chakravorty, A., & Rong, C. (2019). Evaluating Feature Selection Methods for Short-Term Load Forecasting. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp). Presented at the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–8. https://doi.org/10.1109/BIGCOMP.2019.8679188.

Qiong Li, Peng Ren, & Qinglin Meng (2010). Prediction model of annual energy consumption of residential buildings. In *2010 International Conference on Advances in Energy Engineering. Presented at the 2010 International Conference on Advances in Energy Engineering*, pp. 223–226. https://doi.org/10.1109/ICAEE.2010.5557576.

Rahul, K., Seth, N., & Kumar, U. D. (2018). Spotting earnings manipulation: Using machine learning for financial fraud detection. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 343–356). Springer.

Ravi, A. (2020). Stacked Generalization for Human Activity Recognition. arXiv: 2009.10312 [cs].

Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). On developing an automatic threshold applied to feature selection ensembles. *Information Fusion, 45*, 227–245. https://doi.org/10.1016/j.inffus.2018.02.007

Shao, M., Wang, X., Bu, Z., Chen, X., & Wang, Y. (2020). Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society, 57*, 102128. https://doi.org/10.1016/j.scs.2020.102128

Shapi, M. K. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment, 5*, 100037. https://doi.org/10.1016/j.dibe.2020.100037

Sharma, J., Giri, C., Granmo, O.-C., & Goodwin, M. (2019). Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP Journal on Information Security, 2019*, 15. https://doi.org/10.1186/s13635-019-0098-y

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM). Presented at the 2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593–596). https://doi.org/10.1109/ICACTM.2019.8776800

Sonkamble, B. A., & Doye, D. D. (2008). An overview of speech recognition system based on the support vector machines. In *2008 International Conference on Computer and Communication Engineering. Presented at the 2008 International Conference on Computer and Communication Engineering* (pp. 768–771). https://doi.org/10.1109/ICCCE.2008.4580709

Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research, 8*.

Sumaiya Thaseen, I., & Aswani Kumar, C. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences, 29*, 462–472. https://doi.org/10.1016/j.jksuci.2015.12.004

Tahmasebi, M. M., Banihashemi, S., & Hassanabadi, M. S. (2011). Assessment of the Variation Impacts of Window on Energy Consumption and Carbon Footprint. In

Procedia Engineering, 2011 International Conference on Green Buildings and Sustainable Cities 21, pp. 820–828. https://doi.org/10.1016/j.proeng.2011.11.2083.

Vorobeychik, Y., & Wallrabenstein, J. R. (2013). Using Machine Learning for Operational Decisions in Adversarial Environments 9.

Wang, C., Li, Y., & Xu, B. (2018). How Many Versions does a Bug Live in? An Empirical Study on Text Features for Bug Lifecycle Prediction. In *Presented at the The 30th International Conference on Software Engineering and Knowledge Engineering* (pp. 415–461).

Wang, J., Chen, X., Zhang, F., Chen, F., & Xin, Y. (2021). Building load forecasting using deep neural network with efficient feature fusion. *Journal of Modern Power Systems and Clean Energy, 9*, 160–169. https://doi.org/10.35833/MPCE.2020.000321

Wang, W., Zmeureanu, R., & Rivard, H. (2005). Applying multi-objective genetic algorithms in green building design optimization. *Building and Environment, 40*, 1512–1525. https://doi.org/10.1016/j.buildenv.2004.11.017

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. https://doi.org/10.1016/j.enbuild.2018.04.008

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*, 1–37. https://doi.org/10.1007/s10115-007-0114-2

Yu, Z., Haghighat, F., Fung, B. C. M., & Yoshino, H. (2010). A decision tree method for building energy demand modeling. *Energy and Buildings, 42*, 1637–1646. https://doi.org/10.1016/j.enbuild.2010.04.006

Zeng, X.-D., Chao, S., & Wong, F. (2010). Optimization of bagging classifiers based on SBCB algorithm. In *2010 International Conference on Machine Learning and Cybernetics. Presented at the 2010 International Conference on Machine Learning and Cybernetics* (pp. 262–267). https://doi.org/10.1109/ICMLC.2010.5581054

Zhang, C., Cao, L., & Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable Cities and Society, 39*, 508–518. https://doi.org/10.1016/j.scs.2018.02.016

Zhang, L., & Wen, J. (2019). A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings, 183*, 428–442. https://doi.org/10.1016/j.enbuild.2018.11.010

Zhang, X., Wang, J., & Gao, Y. (2019). A hybrid short-term electricity price forecasting framework: Cuckoo search-based feature selection with singular spectrum analysis and SVM. *Energy Economics, 81*, 899–913. https://doi.org/10.1016/j.eneco.2019.05.026

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

Zhao, HaiXiang, & Magoulès, F. (2012). Feature selection for predicting building energy consumption based on statistical learning method. *Journal of Algorithms & Computational Technology, 6*, 59–77. https://doi.org/10.1260/1748-3018.6.1.59

Zhong, H., Wang, J., Jia, H., Mu, Y., & Lv, S. (2019). Vector field-based support vector regression for building energy consumption prediction. *Applied Energy, 242*, 403–414. https://doi.org/10.1016/j.apenergy.2019.03.078

Zhou, C., & Chen, X. (2021). Predicting China's energy consumption: Combining machine learning with three-layer decomposition approach. *Energy Reports, 7*, 5086–5099. https://doi.org/10.1016/j.egyr.2021.08.103