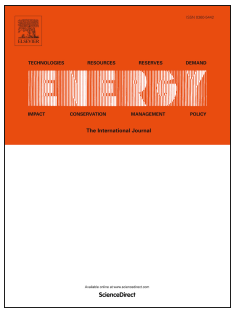
Journal Pre-proof



An interpretable multi-stage forecasting framework for energy consumption and CO_2 emissions for the transportation sector

Hamidreza Eskandari, Qingyao Qiao, Hassan Saadatmand, Mohammad Ali Sahraei

PII: S0360-5442(23)02893-1

DOI: https://doi.org/10.1016/j.energy.2023.129499

Reference: EGY 129499

To appear in: *Energy*

Received Date: 15 July 2023

Revised Date: 31 August 2023

Accepted Date: 27 October 2023

Please cite this article as: Eskandari H, Qiao Q, Saadatmand H, Sahraei MA, An interpretable multi-stage forecasting framework for energy consumption and CO_2 emissions for the transportation sector, *Energy* (2023), doi: https://doi.org/10.1016/j.energy.2023.129499.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Ltd.

Journal Pre-proof

CRediT author statement

Hamidreza Eskandari: Original draft preparation, Writing - Review & Editing, Conceptualization, Methodology, Investigation, Validation, Supervision

Qingyao Qiao: Original draft preparation, Writing - Review & Editing, Software, Visualization, Investigation, Validation

Hassan Saadatmand: Software, Visualization, Validation

Mohammad Ali Sahraei: Original draft preparation

Journal Pre-proof

Journal Pre-proof

An interpretable multi-stage forecasting framework for energy consumption and CO_2 emissions for the transportation sector

Hamidreza Eskandari [a], Qingyao Qiao [b], Hassan Saadatmand [c], Mohammad Ali Sahraei [d]

[a] School of Management, Swansea University, Swansea, United Kingdom

[b] Faculty of Architecture, The University of Hong Kong, Hong Kong

[c] Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[d] Department of Civil Engineering, College of Engineering, University of Buraimi, Oman

Abstract:

The transportation sector is deemed one of the primary sources of energy consumption and greenhouse gases throughout the world. To realise and design sustainable transport, it is imperative to comprehend relationships and evaluate interactions among a set of variables, which may influence transport energy consumption and CO_2 emissions. Unlike recent published papers, this study strives to achieve a balance between machine learning (ML) model accuracy and model interpretability using the Shapley additive explanation (SHAP) method for forecasting the energy consumption and CO_2 emissions in the UK's transportation sector. To this end, this paper proposes an interpretable multi-stage forecasting framework to simultaneously maximise the ML model accuracy and determine the relationship between the predictions and the influential variables by revealing the contribution of each variable to the predictions. For the UK's transportation sector, the experimental results indicate that road carbon intensity is found to be the most contributing variable to both energy consumption and CO_2 emissions predictions. Unlike other studies, population and GDP per capita are found to be uninfluential variables. The proposed multi-stage forecasting framework may assist policymakers in making more informed energy decisions and establishing more accurate investment.

Keywords: Energy consumption forecasting, CO_2 emissions forecasting, Transportation sector, Machine learning, Feature selection.

Journal Pre-proof

1 Introduction

1.1 Background

Over the last decades, energy consumption (EngCons) and carbon dioxide emissions (CO_2E) challenges have been the primary issues for policymakers. This was because of the impact of energy usage on national economic growth as well as the impact of carbon on human health. In parallel with economic and social enhancements, energy demand has risen worldwide. Correspondingly, the rapidly increasing level of the human population, socioeconomic improvement, urbanization, and scientific developments have cumulatively led to an increase in worldwide EngCons and CO_2E in numerous sectors [1, 2]. The world CO_2E for different transportation sectors is provided in Figure 1. Although CO_2E is increasing until 2025, they will continue to decrease dramatically until 2070.

In the United Kingdom (UK), almost all transportation sectors must decarbonize to fulfill the economy-wide net-zero commitment. Due to a continual increase throughout vehicle kilometers traveled, transportation CO_2 peaked in 2007, 8.4% greater compared to 1990. Since then, emissions through the transportation sector have dropped back to around 1990 levels up till 2019, primarily due to enhancements in new vehicle energy efficiency and reduced transportation growth compared to prior years due to a dip following the 2008/2009 recession [3]. According to the Energy Stats [4], although in 2020 the UK government observed significant falls in energy usage for almost all vehicle forms, with the most considerable reduction in buses and automobiles transport remains the most significant part of energy usage throughout the UK.

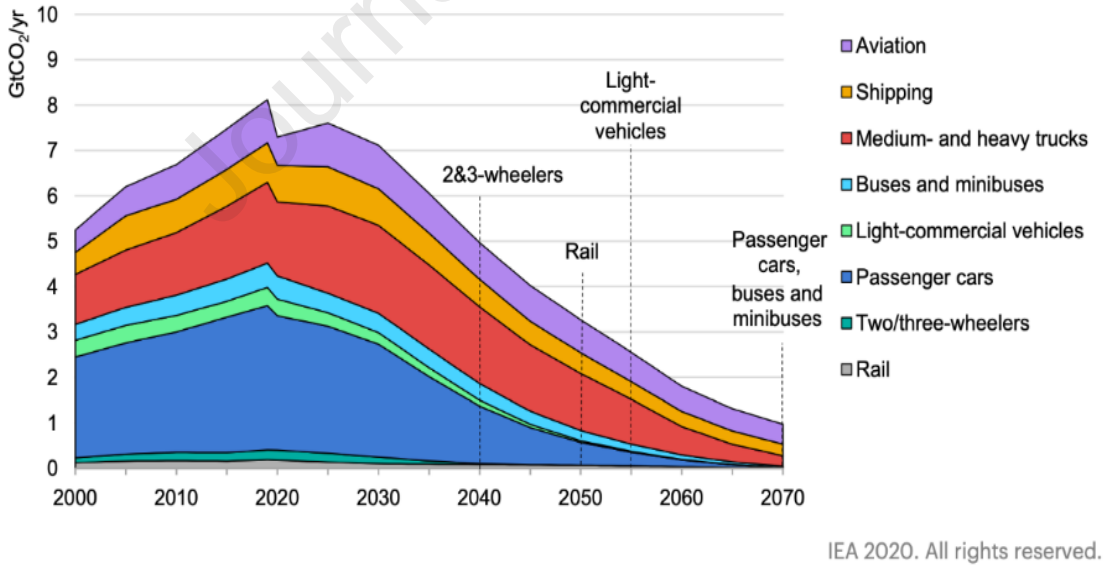


Figure 1. World CO_2E for different transportation modes [5].

A range of machine learning (ML) models can be utilized to estimate EngCons and CO_2E, including multiple linear regression (MLR) [6, 7], logistic regression [8], generalized linear models [9], time series analysis [10, 11], artificial neural networks (ANN) [12-14], deep learning [15-17], support vector machine (SVM) [18, 19], decision tree, random forest (RF) [20, 21], hybrid methods [22, 23], to name

a few. Despite extensive research having been conducted using ML models to forecast EngCons and CO_2E, it is noticed that limited attention has been focused on the transport sector of the UK. For instance, Piecyk and McKinnon [24] conducted a study in 2010 to forecast the carbon footprint of road freight transport in 2020, factors affecting freight transport demand, truck fuel consumption and related CO_2E were discussed. However, their research only focused on part of the transport sector and is now out-of-date, which diminishes the value of their research for policymakers in decision-making. Logan et al. [25] conducted a similar study where they estimated the energy demand of road transport including cars, buses and trains. The energy consumption of the transport sector remained untouched. The limited number of works in energy demand forecasting of the UK's transport sector failed to meet the needs of UK Transport Vision 2050 [26].

Furthermore, as review of literature indicates, previous studies have explored forecasting the EngCons and/or CO_2E, which preselected a small number of features without any strong justification and rationale on how and why those set of features are selected. Because of a small number of preselected features, unlike many studies in other energy forecasting domains, no study has employed FS methods to select the combination of features which can lead to high accuracy of ML models. In addition, only one study [27] has used interpretable ML to forecast the EngCons and CO_2E to determine the influential variables. Therefore, as one of the most crucial needs, the transportation sector is calling for an interpretable multi-stage framework to accurately forecast the EngCons and CO_2E and to systematically evaluate the effects of different types of features.

1.2 Novelty and contributions

The novelty and contributions of this study in comparison with recently published works in forecasting EngCons and CO_2E in transportation sector can be concluded as follows:

- This study considers multi-source data in the UK's transportation sector by integrating three categories of variables (features) including socioeconomic, transportation- and energy-related variables. Neither of published papers used a large list of input features and performed correlation and multicollinearity analyses to remove highly correlated features to provide an appropriate subset of features for interpretability of black-box ML models. All previous works preselected a small number of features without any strong justification and rationale on how and why those set of features are selected.
- Previous studies in the literature used neither filter/embedded methods to select the input features holding strong relationship with the EngCons and CO_2E, nor wrapper methods to select the features which can lead to high accuracy of ML models. This study introduces a novel voting scheme for feature selection (FS), which combines both filter and embedded paradigms, which has not been studied before in the EngCons context.

- The other contribution is to propose an interpretable ML-based forecasting framework, which is depicted in Fig 2. The proposed generalised framework integrates multi-stage FS procedure with ML models to simultaneously achieve more accurate forecasts and more reliable interpretation of black-box ML models using Shapley additive explanation (SHAP) analysis proposed by Lundberg et al. [61]. SHAP is a method employed for interpreting the output of ML models, which is briefly described in Section 3.7. Although the SHAP analysis has recently found applications in energy-related fields [27-29], more work is required to demonstrate the practicality and usefulness of SHAP analysis in interpretable ML models for forecasting EngCons and CO_2E. To the best of our knowledge, this study is the second work in EngCons context (the first study is Aras and Van [30], however, they overlooked the fact that investigation of multicollinearity in their study is crucial) that applies the SHAP analysis to forecast the EngCons and CO_2E to determine the influential variables contributing to the predictive performance.

It is worth mentioning that the proposed methodology is designed for any ML-based forecasting problems, in which simultaneously selecting as few features as possible for interpretability of ML models (or any other reasons) and achieving an acceptable model accuracy are desired. Multiple stages of FS procedure are so insightful on how different features are influencing and interacting with each other.

This paper is organised as follows. Section 2 reviews literature of the existing study regarding EngCons and CO_2E. Section 3 introduces the forecasting framework according to the integrated multi-stage FS methods and ML models. In Section 4, the case study and experimental settings are described. In Section 5, the analyses of experimental results are described along with Shapley Analysis to demonstrate the usefulness and benefits of the proposed framework. Section 6 briefly presents some noticeable discussion and lastly, concluding statements are provided in Section 7.

2 Literature review

2.1 Feature selection for energy forecasting

Multivariate ML-based forecasting models work based on selecting a set of potentially influencing features. As it is discussed in Section 2.2 only Wang et al. [31] used FS method for forecasting CO_2E and EngCons in transportation sector which in contrast to previous studies that have typically chosen a limited number of key features, such as population, GDP, and total number of vehicles, without providing a clear rationale or justification for their selection.

An appropriate FS procedure is crucial in forecasting CO_2E and EngCons. A summary of a few recent studies considering the FS methods in different energy fields is provided as follows. Jurado et al. [32] compared several ML techniques for energy prediction inside houses. They suggested a hybrid strategy that combines FS methods using soft computing and ML models, i.e., fuzzy, RF, and ANN. Feng et al. [33] developed a method for wind prediction using the ML approach. A FS structure was established to identify the most appropriate inputs for the ML approach. An organised FS method for establishing house energy prediction was suggested by Zhang and Wen [34].

The power usage of appliances was forecasted by Moldovan and Slowik [35] utilizing binary grey wolf optimization, in which the best features were selected utilizing the RF, KNN, decision tree, and extra tree methods. Qiao et al. [36] suggested a framework for house energy usage forecasting, depending on FS techniques. Lv and Wang [37] offered an efficient short-term wind speed prediction model by taking into account the impact of several meteorological variables. The filter-wrapper method integrating K-medoid clustering was developed to choose crucial meteorological elements. In the above papers, experimental results indicated that the ML models utilizing the FS methods typically have greater generalization and precision than those without FS.

2.2 ML for energy forecasting

Over the past few years, many forecasting models have been developed to predict EngCons and CO2E in different sectors. In this section, we briefly review the papers which investigated ML-based forecasting in the transportation sector in recent years. Readers can refer to [2,38] to study more references. Wang et al. [39] developed an ML model for transportation emissions utilizing the SVM, GPR, and ANN algorithms. Sahraei et al. [40] forecasted transportation energy usage utilizing the multivariate adaptive regression splines (MARS) method for 45 years after 1975 in Turkey.

Li et al. [41] utilised an ML model to predict Australia's vehicle gasoline usage utilizing an autoregressive and structural method. The outcomes of a prediction regarding gasoline usage for 2019-2020 demonstrate the outstanding prediction performance of the ML model. Ağbulut [2] utilised three ML models, i.e., ANN, SVM, and deep learning to predict transport energy usage and CO2E. Results showed that CO2E and energy usage through the transport sector will rise by almost 3.4 times more by 2050 than today. Sahraei and Çodur [23] suggested hybrid methods, ANN-PSO, ANN-Simulated Annealing, and ANN-GA, for a precise optimization of the input parameters regarding forecasting the energy usage during 1975-2019 throughout Turkey.

Three ML algorithms were developed by Li et al. [42], including gradient boosting regression (GBR), SVM, and ordinary least squares regression, to predict transport CO_2E. More recently, Javanmard et al. [38] employed a hybrid approach integrating a multi-objective mathematical model with MLs to predict energy demand and CO_2E in the transportation sector of Canada. Korkmaz [43] developed black widow optimization and bezier search differential evolution methods to calculate the transport energy usage throughout Turkey. More recently, a hybrid RF-SVR and response surface method were carried out by Khajavi and Rastgoo [44] to forecast CO_2E for 30 important towns throughout China.

Given the concentration of this study, Table 1 summarises key information from prior studies from all over the world forecasting energy demand and CO_2E in the transportation sector. In the case of the FS methods (wrapper, filter, and embedded), we could not find any study to utilise the FS methods except for Wang et al. [31], which used stepwise linear regression to select most significant variables. In addition, based on authors' review of literature only [2,38] considered both energy demand and CO_2E as target variables.

Table1. Summary of the primary studies in predicting energy demand and CO_2E in the transportation sector.

Paper	Target Variable(s)	Field	Country	Timeframe	Feature selection		Applied Models		
					Filter/ Embedded	Wrapper	ML	Mathematical	Hybrid
[2]	Energy and CO_2	Transport	Turkey	1970-2016	No	No	ANN, SVM, DL	-	-
[22]	Energy	Transport	Jordan	1985-2009	No	No	-	-	Neuro-Fuzzy Inference System
[23]	Energy	Transport	Turkey	1975-2019	No	No	-	-	ANN-GA, ANN-SA, and ANN-PSO
[31]	CO_2	Transport	China	1980-2014	Yes	No	BPNN, GPR, SVM	-	PSO-SVM
[45]	Energy	Transport	Thailand	1989-2008	No	No	ANN	-	-
[40]	Energy	Transport	Turkey	1975-2019	No	No	multivariate adaptive regression splines	-	-
[41]	Energy	Automobile	Australia	1974-2019 (Quarterly)	No	No	Autoregressive and structural model	-	-
[42]	CO_2	Transport	Top 30 Emitting Nations	2005-2014	No	No	OLS, SVM, GBR	-	-
[38]	Energy and CO_2	Transport	Canada	1990-2019	No	No	ARIMA, ARFIMA, SARIMA, GARCH, MIDAS, SVR	-	-
[43]	Energy	Transport	Turkey	2000-2017	No	No	Bezier search differential evolution and black widow optimization	-	-
[44]	CO_2	Road Transport	China	2006-2015	No	No	-	-	Hybrid RF-SVR, and response surface
[46]	CO_2	Road Transport	UK	2010-2014	No	No	-	Basic Estimations to 2050	-
[47]	non-methane	Road Transport	EU Nations	2004-2016	No	No	SVR, RLR	-	Kernel grey model

Journal Pre-proof

[48**]**	CO_2	Transport	Pakistan	1971-2014	No	No	-	Autoregressive distributive lag	-
[49**]**	CO_2	Transport	China	2015 (3 months)	No	No	ANN, gaussian naive bayes, linear and logistic regression, stacked deep belief networks	-	-
[50]	CO_2	Land transport	Cyprus	2010-2016	No	No	-	Environmentally extended input-output analysis	-
[51**]**	Energy	Transport	Turkey	1975-2016	No	No	ANN	-	-
Prseent study	Energy and CO_2	Transport	UK	1990-2019	Yes	Yes	ANN, RF, LSTM, SVM, MLR, LSBoost, GPR	-	-

Journal Pre-proof

3. Research methodology

3.1 Overall procedure of proposed forecasting framework

Figure 2 provides a conceptual framework of our proposed method for forecasting EngCons and CO_2E. The primary goal of the proposed integrated multi-stage FS and ML method is to leverage the benefits of FS to identify the most influential features while mitigating multicollinearity, and to identify a subset of features able to achieve an appropriate balance between accuracy of ML models and their interpretability power.

The architecture of our proposed method involves multiple stages: i) pre-processing operations, (ii) correlation analysis of variables, (iii) filter and embedded FS methods, (iv) multicollinearity analysis, (v) integrated wrapper FS and ML models, (vi) selecting the best ML models with the highest accuracy, and finally (vii) performing Shapley analysis to determine the contributions of variables.

First, the raw data was collected and processed by removing noise, correcting the data inconsistencies, and integrating them into a homogeneous dataset. The other stages are described in some details as follows. In any ML study, some potentially important features can be automatically selected and added to the feature list based on the domain knowledge as depicted in Figure 2.

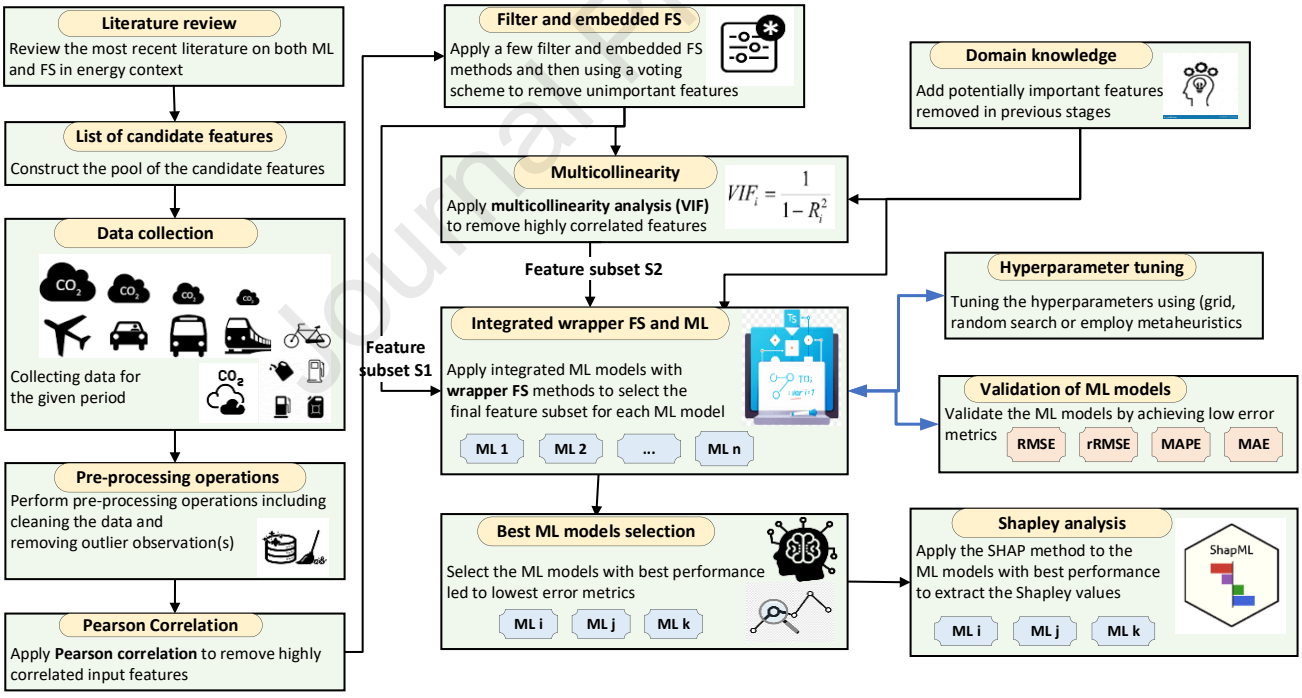


Figure 2. The conceptual framework of the integrated FS and ML based forecasting framework.

3.2 Pairwise Pearson correlation

For each input feature (IF), the Pearson correlation coefficient r_{ij} of this feature (i) with each of the other input features (j) along with its corresponding significance value p_{ij} (i and $j \in Set_{IF}$) are calculated. If the absolute value of correlation coefficient r_{ij} for each pair of input features is greater than correlation

threshold value T_u (in this study T_u = 0.95) with significant confidence ($p_{ij} < a$, a is significance level), only one highly collinear feature will be selected, while the remaining features will be excluded from the candidate pool to avoid redundancy.

3.3 Filter and embedded FS methods

3.3.1 Maximum Relevance and Minimum Redundancy (mRMR)

The mRMR is a filter FS algorithm to rank input features based on their relevance to the output feature and simultaneously discard redundant input features [52]. Mutual information (MI) is employed to quantify both the relevance and redundancy of mRMR. The following describes MI:

$$I(X,Y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where X, Y are vectors, $p(x,y)$ is the joint probabilistic density, $p(x)$ and $p(y)$ are the marginal probabilistic densities, respectively.

Assuming a feature set S with m ($x_i,\ i \in (1,m)$) features, then Max-Relevance represents a feature subset that jointly has the largest relevance to the output variable y:

$$maxD(S,y),\ D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i,y) \quad (2)$$

For defining redundant features, Minimum Redundancy is implemented using Max-Relevance's possible redundancy. :

$$minR(S),\ R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i,,x_j) \quad (3)$$

An incremental search method is then employed to find the optimal solution that can satisfy the above two constraints. Assuming that there already have a feature set S_{m-1} , the task is to determine the mth feature from $\{X - S_{m-1}\}$

$$max_{x_j \in X - S_{m-1}}[I(x_j,y) - \frac{1}{m-1}\sum_{x_i \in S_{m-1}} I(x_i,x_j)] \quad (4)$$

3.3.2 Random Forest

Random Forest (RF) is an ensemble technique that combines a predetermined number of decision trees. It employs information gain or Gini impurity as the criteria for splitting each node across all trees [53]. Nodes with the highest impurity reduction are usually found at the start of decision trees, while nodes

with the lowest impurity reduction are typically located towards the end. Therefore, by selectively pruning branches at a specific node, it is feasible to create a subset of the most significant features.

3.3.3 Boruta FS

Boruta is a wrapper algorithm that utilises an RF to identify pertinent features associated with output labels while discarding irrelevant features that may occasionally exhibit significance due to chance [54]. A detailed procedure for BFS is iterated below:

- Randomise the feature set by creating shadow copies (shadow features) of all features and combining them with the original features to create an extended feature set.
- Establish an RF model for the expanded feature set and assess the feature's importance (the average reduced accuracy Z value). The larger the Z value, the more significant the trait. Z-max denotes the maximum Z value of the shadow feature.
- During each iteration, if the feature's Z value is larger than Z-max, the feature is deemed essential and is retained. Otherwise, the feature will be considered trivial and eliminated from the feature list.
- The preceding approach terminates when all features are either verified or rejected or when the maximum number of BFS iterations is achieved.

3.3.4 Voting scheme in FS

One of the challenges in FS is to determine the most appropriate FS method(s) for a particular set of data due to the fact that each FS method has its own logic based on a statistical measure to calculate the relative importance of features, which may lead to different subsets of selected features. In other words, a feature may be deemed important in one method but not in another.

In the proposed methodology, three different FS methods, including two filter FS methods (mRMR and Boruta) and one embedded FS method (RF) were applied to select the most important features. However, both filter and embedded methods have their own drawbacks; filter methods ignore the dependency among input features and embedded methods heavily rely on ML models. To overcome such issue, a voting scheme was proposed, when an FS method *m* picks the feature *f*, it assigns the vote V_{mf} (*no* = 0 or *yes* = 1) for that feature. In the end, the voting scheme calculates the total "*yes*" votes for each feature, i.e., total vote $V_f = \sum_m V_{fm}$, and then a subset of features with the total vote V_f greater than or equal to the total vote threshold value V_T (provided by ML practitioner depending on the problem) are selected.

3.4 Multicollinearity analysis

Multicollinearity refers to the correlation between input features in ML applications, which typically does not impact the performance of the ML models. However, it may significantly distort the interpretability of the model and develop a biased insight of feature importance. For instance, during

training process, a ML method such as Lasso Regression may assign a large weight to one arbitrary representative of a group of highly correlated features and fully omit the rest ones, despite similar information these features represent. A misleading interpretation may therefore be obtained due to multicollinearity. To mitigate adverse impact of multicollinearity on interpretation of ML, variance inflation factor (VIF) [55, 56] was introduced in the study for removing highly correlated features.

3.5 ML models

3.5.1 Support vector machine (SVM)

SVM is a model for binary classification that functions based on the principle of separating hyperplanes [57]. This approach guarantees the determination of the hyperplane that can effectively partition the training datasets under the largest geometric interval. A Kernel function is incorporated into SVM to facilitate the mapping of input spaces onto a feature space of high dimensionality through a non-linear transformation, which ultimately results in the establishment of a linear decision boundary within the transformed space.

3.5.2 Gaussian process regression (GPR)

The GPR model is a supervised machine-learning algorithm that operates on probabilistic principles [58]. It leverages prior knowledge to generate predictions and provides measures of uncertainty. Assuming a training set $\mathcal{D} = (X, y) = \{(X_i, y_i)|i = 1, \dots, N\}$, where X is denotes an input vector and y denotes an output or target variable. When given a new input X^*, the corresponding output $\hat{y}^*$ can be expressed as

$$\hat{y}^* = K(X^*, X)K(X, X)^{-1}y \tag{5}$$

The derivation process is as follows, assuming:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}) \tag{6}$$

According to the conditional distribution property of the multidimensional Gaussian distribution:

$$y^*|y \sim \mathcal{N}(K(X^*,X)K(X,X)^{-1}y, K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*)) \tag{7}$$

Finally, $p(\hat{y}^*|y)$ is able to achieve its maximum when $\hat{y}^* = K(X^*, X)K(X, X)^{-1}y$.

3.5.3 Long short-term memory (LSTM) networks

Long Short-Term Memory (LSTM) networks are sequential neural networks that address the vanishing gradient of Recurrent Neural Networks [59] by introducing the concept of cell states and bring four interacting layers and gate units as shown in Figure 3.

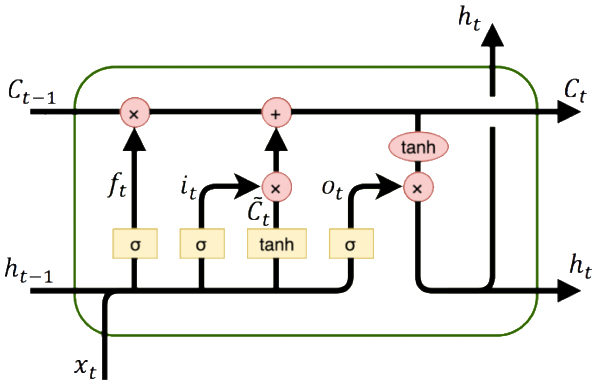


Figure 3. Schematic diagram of an LSTM.

The self-connected memory cell C_t is the key feature of LSTMs, enabling gradients to flow across long sequences. LSTMs use 3 sigmoid gates (forgetting, input, and output) to manage cell state information.:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{8}$$

Forgetting gate f_t determines the specific information to discard from the cell state based on h_{t-1} and x_t, and update the cell state C_{t-1} (i.e., 0: discard, 1: remain).

$$f_t = \sigma\left(W_f \cdot [\, h_{t-1}, x_t] + b_f\right) \tag{9}$$

The input gate i_t and a *tanh* layer are then developed to control new information stored in the new cell:

$$i_t = \sigma(W_i \cdot [\, h_{t-1}, x_t] + b_i) \tag{10}$$
$$\tilde{C}_t = tanh(W_C \cdot [\, h_{t-1}, x_t] + b_C) \tag{11}$$

the new cell state C_t can be updated:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{12}$$

Finally, the output gate o_t uses the current input and the previous output to decide what parts of the cell state to output, and a *tanh* function is established to calculate the current state.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{13}$$
$$h_t = o_t \cdot \tanh(C_t) \tag{14}$$

In Equations (8)-(14), the matrices W_f, W_i and W_o are the recurrent weighting metrics; b_f, b_i, b_C and b_o are the corresponding bias vectors.

3.5.4 Linear regression (LR)

LR measures the relationship between a target variable and a given set of input variables [60]. Assuming there are *m* input variables:

$$Y = \beta_0 + cX_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon \quad (15)$$

where β_0 is the constant term and $\beta_1$to β_m are the coefficients associated with the input variables. ε is the random error. Note that the m^{th}regression coefficient β_m represents the expected change in Y per unit change in the m^{th} input variable x_m, assuming $E(\varepsilon) = 0$, $\beta_m = \frac{\partial E(Y)}{\partial X_m}$.

3.5.5 Gradient tree boosting with least squares (LSBoost)

LSBoost is a meta learning method that comprises a specific number of weak tree-learners [61].The algorithm initiates by sequentially training individual weak learners in the form of decision trees, and subsequently fits the residual of errors to attain improved performance. The LSBoost approach employs the least squares as the criterion for loss.

Assuming the training set $\{(x_i, y_i)\}_{i=1}^{n}$, a loss function $L(y, F) = \frac{(y-F)^2}{2}$ and regression function $F_m(x)$, $(F_0(x) = \bar{y})$, where m is the number of iterations.

For each iteration,

$$\tilde{y_i} = y_i - F_{m-1}(x_i), \text{for i} = 1,2, \ldots, \text{N} \quad (16)$$

$$(\rho_m, \alpha_m) = argmin_{\rho,\alpha} \sum_{i=1}^{N} [\tilde{y_i} - \rho h(x_i; \alpha)]^2 \quad (17)$$

$$F_m(x) = F_{m-1}(x) + \rho_m h((\text{x}; \alpha_m) \quad (18)$$

where $h(x_i; \alpha)$ is a parameterised function of input variables x_i that characterised by parameter α_m, ρ_m is a successive increment/step/boost of LSBoost.

3.5.6 Multi-layer Perceptron (MLP)

A multilayer perceptron is a fully connected feedforward artificial neural network (ANN) containing at least three layers of nodes: an input layer, a hidden layer, and an output layer [62].

Assuming an input layer consisting of a set of neurons $\{x_i | x_1, x_2, \ldots, x_m\}$, each neuron in the hidden layer is linearly weighted to sum the values from the input layer:

$$v_i = \omega_{i1} x_1 + \omega_{i2} x_2 + \cdots + \omega_{im} x_m \quad (19)$$

Where v_i is the weighted sum of the input connections of hidden node i , ω_{im} is the weight between hidden node i and input x_m.

Then, the weighted summation is applied to a nonlinear activation function, typically a hyperbolic tan function or sigmoid function:

$$y(v_i) = \tanh(v_i) \quad \text{or} \quad y(v_i) = (1 + e^{-v_i})^{-1} \tag{20}$$

The learning process in the MPL is carried out through backpropagation by changing the weights after all data is processed.

Assuming an error in an output node j in the nth data point $e_j(n) = y_j(n) - \bar{y}_j(n)$, where y is the actual value and $\bar{y}$ is the calculated value. The node weights can be adjusted based on the least mean squares algorithm to minimise the error in the entire output as described:

$$\mathcal{E}(n) = \frac{1}{2}\sum_j {e_j}^2(n) \tag{21}$$

According to gradient descent, the change in each weight is:

$$\triangle\, \omega_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} \bar{y}_j(n) \tag{22}$$

Where $\bar{y}_j$ is the output of the previous neuron and η is the learning rate, then the derivative can be described with Equation (25):

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'\left(v_j(n)\right)\sum_k -\frac{\partial \mathcal{E}(n)}{\partial v_k(n)}\omega_{jk}(n) \tag{23}$$

Where ϕ' is the derivative of the activation function. The derivative depends on the change in weights of the kth nodes, which represent the output layer.

3.6 Wrapper FS

SIFE [63] is an efficient wrapper-based evolutionary algorithm with set-inspired operations and fuzzy granulation for both high-dimensional and low-dimensional FS problems. To improve its search policy, SIFE uses a three-parent crossover approach based on set-theoretic concepts such as ‘union’ and ‘intersection’. Fuzzy granulation is also integrated into SIFE, which aids in population initialization and elite steps. It helps in generating a diverse population throughout generations and acts as a surrogate strategy to avoid additional fitness evaluations. This approach aims to achieve a fast and reasonable balance between exploration and exploitation in its problem encoding and search operation, while

reducing computational costs. SIFE is adopted in this research because of its high capability of handling both high-dimensional and low-dimensional search space.

3.7 ML interpretation by SHAP method

As black-box ML models have long been criticised for lacking interpretability. In the context of decision-making, stakeholders and policy makers prioritise quantitative analysis of the correlation between input and target variables over the predictive accuracy of ML models.

Shapley Additive Explanations (SHAP) is an approach employed for interpreting the output of ML models [64]. The classical Shapley value from game theory is used by SHAP method to establish a connection between the optimal credit allocation and the local explanation. SHAP operates by decomposing the output of an ML model into the sums of the impacts of individual features which facilitates the comprehension of the significance of individual features and benefits decision-making. In order to compute SHAP value, a linear explanation model is utilised as an interpretable approximation to a ML model [65]:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{24}$$

where $z' \in \{0,1\}^M$ represents whether a feature is used to estimate the output variable, M is the number of input features, ϕ_i is the SHAP value of the ith feature, and ϕ_0 is the mean value of the output variable. The SHAP value assesses feature importance by comparing model prediction performance with and without each feature in feature combinations:

$$\phi_i = \sum_{S \subseteq z' \{i\}} \frac{|S|!(M-|S|-1)}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{25}$$

where S is the set of non-zero z', and $f_x(S) = E[|f(x)x_S|]$ is the expected outcome of the model $f(x)$ subjected to S.

4 Experimental setting

The study utilises the MATLAB software to implement the ML algorithms and wrapper FS. All experiments for filter and embedded FS, SHAP analysis and producing heatmap are coded in Python.

4.1 Data

The case of this study is based on the EngCons and CO_2E in the UK's transportation sector in the time interval of 1990-2019. The year 2020 was not considered in this study, because this observation is detected as an outlier for some features due to the Covid Pandemic. The EngCons and CO_2E of the UK's transportation sector is illustrated in Figure 4, where similar pattern as bimodal distribution chart

in EngCons and CO_2E is observed. The initial growth trajectory encountered an avalanche of decline in 2008 which persisted until 2013 when it reached its lowest point, the same as in 1990. After a brief four-year increase in EngCons and CO_2E levels, they began a second decline in 2017 and beyond.

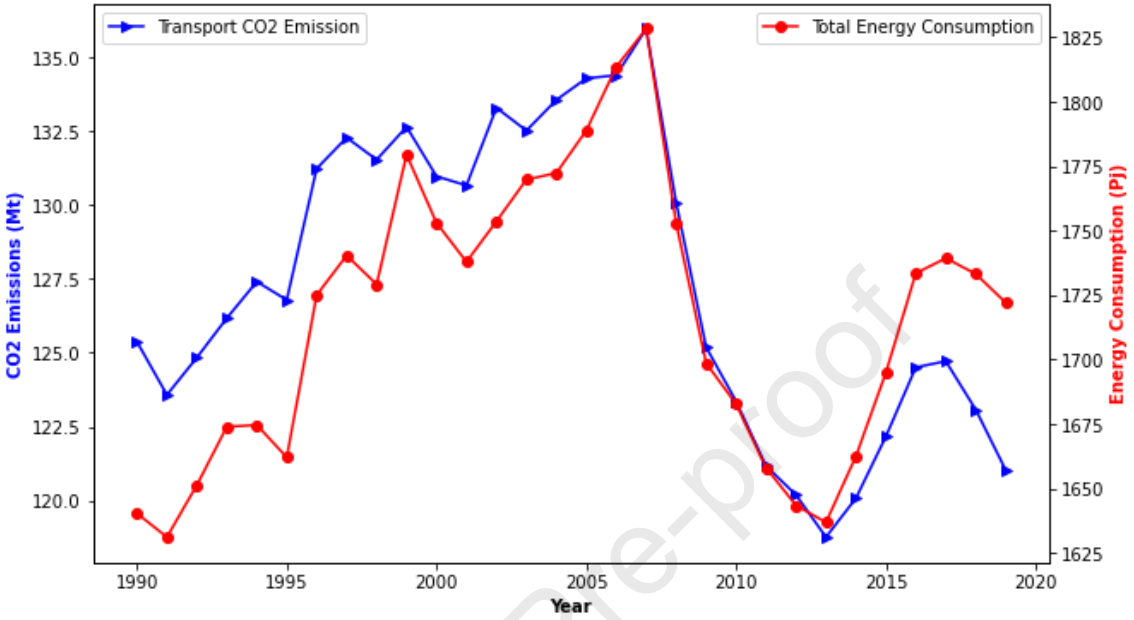


Figure 4. EngCons and CO_2E in the transportation sector in the UK (1990-2019).

Table 2 presents the list of 24 variables (features) along with their corresponding abbreviations, which include 22 input variables and two target variables, EngCons and CO_2. Based on the intensive review of existing literature, a wide range of input variables in three categories including socioeconomic, transportation and energy-related factors are considered. Socioeconomic factors include GDP per capita, population, gasoline price and unemployment rate which indicate the strong relationship between the EngCons and CO_2E in literature. To consider urbanization level of the UK, urban population rate is added to the list.

For transportation category, we considered three main modes of transportation including air, rail, and road transportation, and each mode comprising passenger and freight transportation. Furthermore, energy intensity in transport, renewable and waste energy in transport, share of electric vehicles and road carbon intensity which could have significant influence on EngCons and CO_2E are considered. For energy category, a few energy-related variables considered in other studies, which may have relationships with EngCons and CO_2E in transport are added to the list.

The data was collected from the UK Department for Transport (www.gov.uk/government/organisations/department-for-transport), UK Office for National Statistics (www.ons.gov.uk/economy/environmentalaccounts), and International Energy Agency (World Energy Balances Highlights) https://www.iea.org/data-and-statistics/data-product/world-energy-balances-highlights. The descriptive statistics of the 24 features considered for transport EngCons and CO_2E is presented in Table A1.

Table 2. Twenty-four features and their corresponding abbreviation.

Feature	Abbreviation	Feature	Abbreviation
CO_2E in transport	CO_2	Unemployment rate	UR
EngCons in transport	EC-Trans	Gasoline price	GP
Oil Products consumption in transport	OP-Trans	Total public energy RD&D budget	R&D
Total EngCons in all sectors	EC-All	Net energy imports	NEI
Total energy supply	TES	Road carbon intensity	RCI
Population	POP	Air passengers	AP
Urban population rate	UPR	Air fright	AF
Energy intensity in transport	EI	Rail passengers	RP
GDP per capita	GDP	New road vehicle registrations	NVR
Total final Electricity consumption	EC-Elec	Total licensed road vehicles	TV
Renewable & Waste by Transport	RW-Trans	Average road vehicle milage	AVM
Share of electric vehicles	SEV	Total road vehicle milage	TVM

4.2 Hyperparameter setting of MLs

Optimising hyperparameters is crucial for MLs resulting in promising performance. A pilot study was conducted to manually tune the parameters to assess the impact of FS methods on ML model performance and evaluate the influence of different features on EngCons and CO_2E. The best values of hyperparameters found in the pilot experiments for seven ML models are presented in Table 3.

Table 3. The summary of hyperparameters for the ML models.

ML Model	Parameter name	Parameter value
MLR	-	-
RF	Maximum number splits	10
	Pruning	Off
	Number of trees	7
LS-Boost	Maximum number splits	12
	Pruning	Off
	Number of trees	7
	Learning rate	0.3
MLP	Number of hidden layers	1
	Number of neurons	10
	Learning method	Levenberg-Marquardt
	Activation function	Sigmoid
LSTM	Number of LSTM layers	1
	Number of hidden units	20
	Maximum epoch	100
SVR-RBF	Box constraint I	100
	Gamma (γ)	1
GPR	Explicit basis	Linear
	Prediction method	Subset of regressors approximation

4.3 Evaluation metrics

To develop a comprehension of the model performance in terms of the EngCons and CO2E of transportation sector, the following evaluation metrics are included: root mean square error (*RMSE*), relative root mean square error (*rRMSE*), mean absolute error (*MAE*), mean absolute percentage error (*MAPE*).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2} \tag{26}$$

$$rRMSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}{\sum_{i=1}^{n}(p_i)^2}} \tag{27}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i| \tag{28}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - p_i}{y_i}\right| \tag{29}$$

where y_i is the actual value and p_i is the predicted value.

5. Results and analysis

5.1 Primary correlation analysis

Since all 22 input variables are continuous numerical, Pearson correlation coefficients with two-tailed test were computed among them. The correlation results indicate that some correlation values were both statistically significant with a = .01 and greater than predefined correlation threshold .95.

The correlation between population and urban population rate was found to be extremely high, $r(28)$=+1. The variables rail passengers and air passengers were found to be extremely positively correlated with population with $r(28)$= .99 and $r(28)$= .98, respectively. Energy intensity is extremely negatively correlated with population, $r(28)$= -.98. It is interesting to note that gasoline price has very strong positive correlation with GDP per capita, $r(28)$ = .93. Total number of licensed road vehicles has very strong positive correlation with population, $r(28)$ = .95, but our domain knowledge recommended us not to remove TV.

Correlation analysis of target variables indicated that EC-Trans(t) and CO2(t) have very strong positive correlation of $r(28)$= .84. Renewable and waste energy in transport and share of electric vehicles have slight or very weak negative correlation with EC-Trans(t), but moderate negative correlation with CO2(t). The correlation values of EC-Trans(t) and CO2(t) with their corresponding next year values, i.e. EC-Trans(t+1) and CO2(t+1), are $r(27)$= .86 and $r(27)$= .91, respectively.

5.2 Intermediate results

After removing extremely correlated input features, voting scheme was implemented in a conservative manner with total vote threshold value V_T of 1. In fact, features were discarded only if they be deemed

not important in all 3 FS methods. The selected feature subsets for EC-Trans(t+1) and CO2(t+1) were listed in Tables 4 and 5.

In this study, regarding that population and GDP per capita variables are found to be important in all similar studies, domain knowledge recommended adding these two variables to the selected list of features after they were removed in the voting scheme.

Table 4. Result of FS voting scheme for EC-Trans(t+1).

FS Method	CO2(t)	EC-Trans(t)	OP-Trans	EC-All	TES	R&D	UR	RCI	NVR	TV	Total count
mRMR	●	●	●	●	○	○	●	●	●	●	8
Boruta	●	●	●	●	●	○	●	●	●	○	8
RF	●	●	●	●	○	●	●	●	●	○	8
Score	3	3	3	3	1	1	3	3	3	1	10

Table 5. Result of FS voting scheme for CO2(t+1).

FS Method	CO2(t)	EC-Trans(t)	OP-Trans	EC-All	TES	RW-Trans	R&D	NEI	RCI	AVM	GP	POP	Total count
mRMR	●	●	●	●	●	●	●	●	●	○	○	●	10
Boruta	●	○	●	●	●	●	●	●	●	●	●	○	10
RF	●	●	●	●	●	●	●	●	●	●	○	○	10
Score	3	2	3	3	3	3	3	3	3	2	1	1	12

The correlations between selected featurses and EC-Trans(t) and CO2(t) were presented in Figures 5 and 6 where significant correlative features were detected, which confirms the need for multicollinearity analysis.

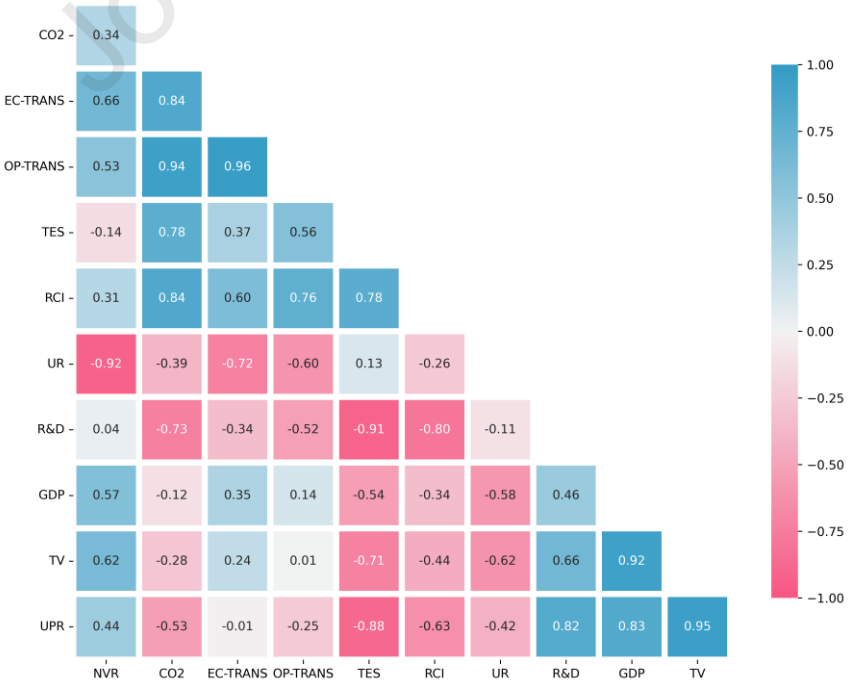


Figure 5. The heatmap of correlation coefficient for EC-Trans(t).

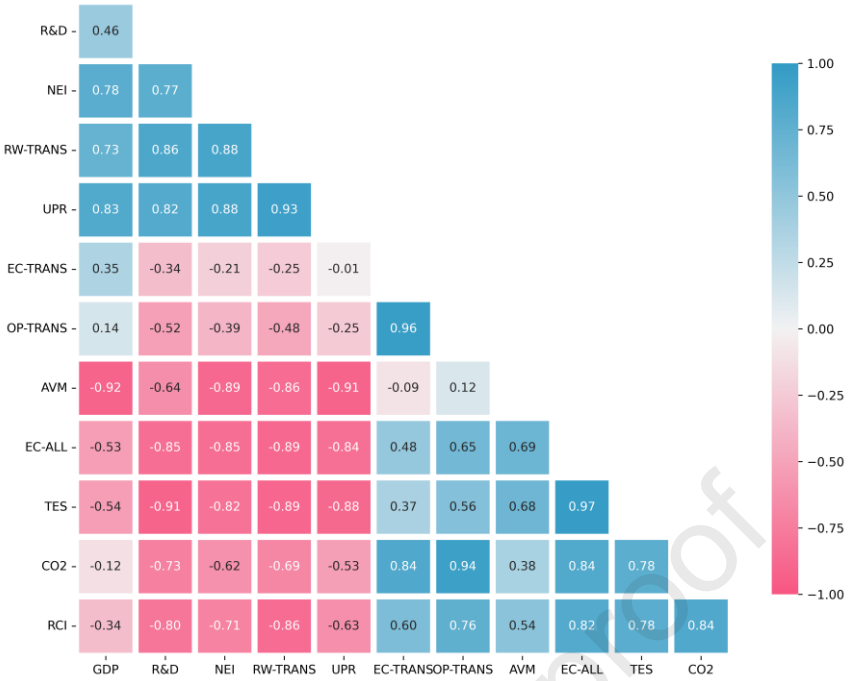


Figure 6. The heatmap of correlation coefficient for CO2(t).

5.3 Multicollinearity analysis

Multicollinearity analysis was performed to model the relationship between 12 selected features and EC-Trans(t+1), and between 13 selected features and CO2(t+1) as presented in Tables 4 and 5, respectively. The Multicollinearity analyses indicate that two regression models have severe multicollinearity for some of the features. For both EC-Trans(t+1) and CO2(t+1) regression models, the features with the highest VIF are iteratively removed until the VIF for each feature becomes less than 10. In EC-Trans(t+1) regression model, the OP-Trans, CO2(t), TES, POP and TV were respectively removed, whereas in CO2(t+1) regression model, OP-Trans, R&W-Trans, TES, POP, AVM and GDP were respectively removed.

Each selected feature subset, before and after multicollinearity analysis (called S1 and S2, respectively, as shown in Figure 2), underwent the final wrapper FS method (SIFE) to produce the final feature subsets as listed in Tables 6 and 7. More specific, for both EC-Trans(t+1) and CO2(t+1), the number of selected features by ML models based on S1 varied significantly while constantly for subset S2. On the other hand, intersections in FS were noticed despite individual preference of ML models in determining the crucial features. For EC-Trans(t+1), NVR and TV are extra two features deemed as the most influential features by all ML models in S1 followed by GDP. In S2, NVR and OP-Trans are determined as important features. For CO2(t+1), in S1, AVM, GP, and RCI was determined significant for 7, 6 and 5 times, respectively. When it comes to S2, 4 features including GP, RCI, EC-Trans and OP-Trans were selected multiple times (i.e., 7, 7, 6 and 6 times) by ML models.

It's important to note that unlike similar studies where population and GDP are key driving variables, in the UK's transportation sector only GDP is selected as influential variable for forecasting EC-

Trans(t+1). The possible reason for this observation could be due to fact that although in all countries, particularly developing countries, population and GDP are key driving factors of EngCon and CO2E, these two measures are significantly harnessed in the UK and they have been decreasing since 2017 as depicted in Figure 4.

Table 6. Selected feature subsets with the ML models for forecasting EC-Trans(t+1).

Feature subset	ML models	Selected features												Number of features
		CO2(t)	OP-Trans	EC-All	TES	R&D	UR	RCI	NVR	TV	POP	GDP	EC-Trans(t)	
S1	LSBoost	○	●	●	○	●	○	○	●	●	○	●	●	7
	MPL	●	●	●	○	○	○	○	●	●	○	●	●	7
	SVR	○	○	○	○	○	○	●	●	●	●	○	●	5
	GPR	○	○	○	○	○	●	○	●	●	○	○	●	4
	LR	○	○	○	○	○	●	●	●	●	○	●	●	6
	RF	○	●	●	●	●	●	●	●	●	●	●	●	11
	LSTM	○	●	○	○	○	○	●	●	●	○	●	●	6
	Score	1	4	3	1	2	3	4	7	7	2	5	7	46
S2	LSBoost	N/A	N/A	●	N/A	●	●	○	●	N/A	N/A	○	●	5
	MPL			●		○	●	○	●			●	●	5
	SVR			●		○	○	○	●			●	●	4
	GPR			●		○	●	○	●			○	●	4
	LR			●		○	●	●	●			○	●	5
	RF			●		○	○	●	●			●	●	5
	LSTM			○		○	○	●	●			●	●	5
	Score			6		1	4	3	7			4	7	32

Note: N/A indicates the feature is excluded in S2

Table 7. Selected feature subsets with the ML models for forecasting CO_2(t+1).

Feature subset	ML models	Selected features													Number of features
		EC-Trans(t)	OP-Trans	EC-All	TES	RW-Trans	R&D	NEI	RCI	AVM	GP	POP	GDP	CO2(t)	
S1	LSBoost	○	●	○	○	●	○	○	○	●	○	○	○	●	4
	MPL	●	○	○	○	○	●	○	●	●	●	○	●	●	7
	SVR	○	○	●	○	○	○	●	●	●	●	○	○	●	6
	GPR	○	○	○	○	○	●	○	○	●	●	○	○	●	4
	LR	●	○	●	○	○	●	○	●	●	●	○	●	●	8
	RF	○	●	●	○	●	○	○	●	●	●	○	○	●	7
	LSTM	●	●	○	●	○	●	●	●	●	●	●	●	●	11
	Score	3	3	3	1	2	4	2	5	7	6	1	3	7	47
S2	LSBoost	●	N/A	●	N/A	N/A	○	○	●	N/A	●	N/A	N/A	●	5
	MPL	●		●			○	●	●		●			●	6
	SVR	○		●			○	●	●		●			●	5
	GPR	●		○			●	●	●		●			●	6
	LR	●		●			●	○	●		●			●	6
	RF	●		●			●	○	●		●			●	6
	LSTM	●		●			○	●	●		●			●	6
	Score	6		6			3	4	7		7			7	40

Note: N/A indicates the feature is excluded in S2

5.4 Final results

The performance of ML models in terms of *RMSE, rRMSE, MAPE* and *MAE* metrics for forecasting EC-Trans(t+1) and CO2(*t*+1) based on S1 and S2 feature subsets are listed in Tables 8 and 9 as well as

Figures 8 and 9, respectively. As shown in the tables and figures, the performance metrics share a similar pattern in both EC-Trans(t+1) and CO2(*t*+1) forecasts. In general, a better performance was observed when the ML models were treated with more features. The performance of some ML models was relatively constant or even improved with a reduced number of features. For example, GPR indicated an increase in all metrics in EC-Trans(t+1) despite fewer input features. Similar results were shown in SVR-RBF and LSTM regarding CO_2E prediction. Among all employed ML models, SVR-RBF and RF indicated the best performance throughout all feature subsets and prediction tasks even with fewer features compared with other ML models.

Table 8. ML performance for forecasting EC-Trans(t+1) with S1 and S2 feature sets.

Dataset	**ML models**	**Metrics**				**Nubmer of features**
		RMSE	rRMSE	MAPE	MAE	
S1	LSBoost	14.805	0.864	0.806	13.770	7
	MLP	17.160	1.000	0.864	14.903	7
	SVR-RBF	6.769	0.395	0.338	5.833	5
	GPR	21.131	1.233	1.148	19.647	4
	LR	13.847	0.808	0.633	10.876	6
	RF	10.881	0.635	0.488	8.404	11
	LSTM	38.311	2.235	2.129	36.575	6
Average		17.558	1.024	0.915	15.715	6.571
S2	LSBoost	23.668	1.381	1.092	18.763	7
	MLP	33.500	1.954	1.569	27.036	5
	SVR-RBF	8.357	0.488	0.340	5.880	4
	GPR	19.561	1.141	0.849	14.666	4
	LR	26.552	1.549	1.375	23.675	5
	RF	17.159	1.000	0.702	12.079	5
	LSTM	40.154	2.342	2.248	38.649	6
Average		24.136	1.408	1.168	20.109	5.143
Difference	LSBoost	59.9%	59.8%	35.5%	36.3%	0.0%
	MLP	95.2%	95.4%	81.6%	81.4%	-28.6%
	SVR-RBF	23.5%	23.5%	0.6%	0.8%	-20%
	GPR	-7.4%	-7.5%	-26.0%	-25.4%	0.0%
	LR	91.8%	91.7%	117.2%	117.7%	-16.7%
	RF	57.7%	57.5%	43.9%	43.7%	-54.5%
	LSTM	4.8%	4.8%	5.6%	5.7%	0.0%
Average		37.5%	37.4%	27.6%	27.9%	-21.7%

Table 9. ML performance for forecasting CO2(*t*+1) with S1 and S2 feature sets.

Dataset	**ML models**	**Metrics**				**Number of features**
		RMSE	rRMSE	MAPE	MAE	

	LSBoost	1.180	0.962	0.876	1.072	4
	MLP	1.011	0.825	0.693	0.849	7
	SVR-RBF	1.16	0.946	0.734	0.902	6
S1	GPR	1.197	0.977	0.867	1.068	4
	LR	1.426	1.163	0.837	1.031	8
	RF	1.311	1.069	0.976	1.197	7
	LSTM	1.920	1.566	1.368	1.682	11
Average		1.315	1.073	0.907	1.114	6.714
	LSBoost	1.412	1.152	0.919	1.132	5
	MLP	1.901	1.554	1.377	1.690	6
	SVR-RBF	1.126	0.918	0.751	0.925	5
S2	GPR	2.041	1.665	1.378	1.688	6
	LR	1.484	1.211	0.876	1.079	6
	RF	1.867	1.523	1.203	1.483	6
	LSTM	1.787	1.458	1.278	1.570	6
Average		1.660	1.354	1.112	1.367	5.714
	LSBoost	19.7%	19.8%	4.9%	5.6%	25.0%
	MLP	88.0%	88.4%	98.7%	99.1%	-14.3%
	SVR-RBF	-2.9%	-3.0%	2.3%	2.5%	-16.7%
Difference	GPR	70.5%	70.4%	58.9%	58.1%	50.0%
	LR	4.1%	4.1%	4.7%	4.7%	-25.0%
	RF	42.4%	42.5%	23.3%	23.9%	-14.3%
	LSTM	-6.9%	-6.9%	-6.6%	-6.7%	-45.5%
Average		26.2%	26.3%	22.5%	22.6%	-14.9%

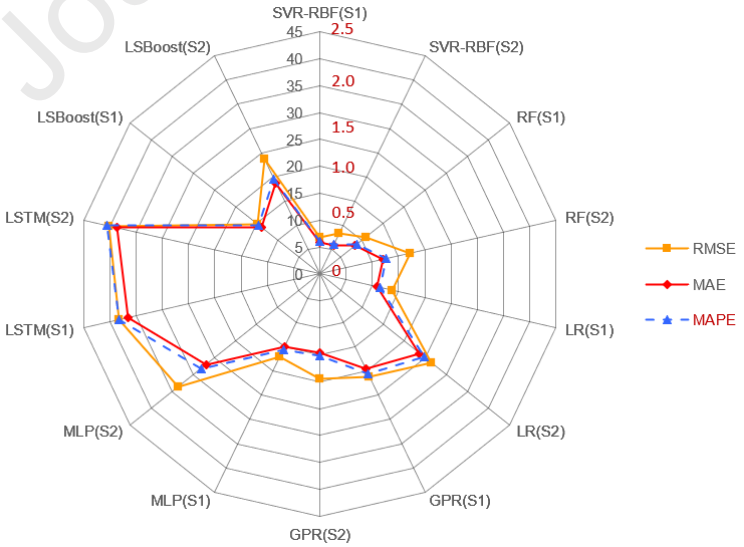


Figure 7. Performance of the ML models in forecasting EC-Trans(t+1).

Journal Pre-proof

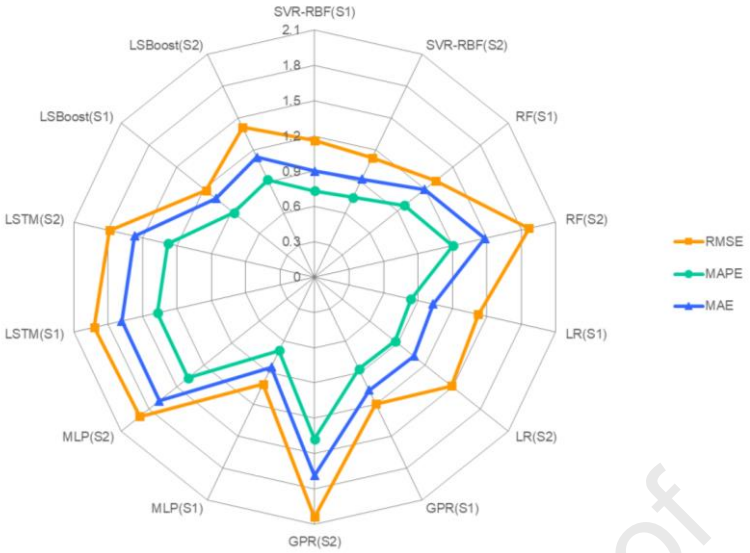


Figure 8. Performance of the ML models in forecasting CO2(t+1).

Figures 10 and 11 illustrate the detailed training and testing performances of SVR-RBF using feature subsets S1 and S2 for forecasting EC-Trans(t+1) and CO2(t+1), respectively. It is noticed that there was only a marginal decrease in *rRMSE* when utilising a smaller feature subset. Specifically, S2 had one less feature than S1 for both EngCons and CO_2E. This result indicates that the FS procedure proposed in the study was successful in reducing the dimension of the feature set while preserving the most pertinent information. In addtion, the evaluation of SVR-RBF on the test set exhibits a negligible diminution in its performance as compared to that on the training set, which indicates slight overfitting issue.

Journal Pre-proof

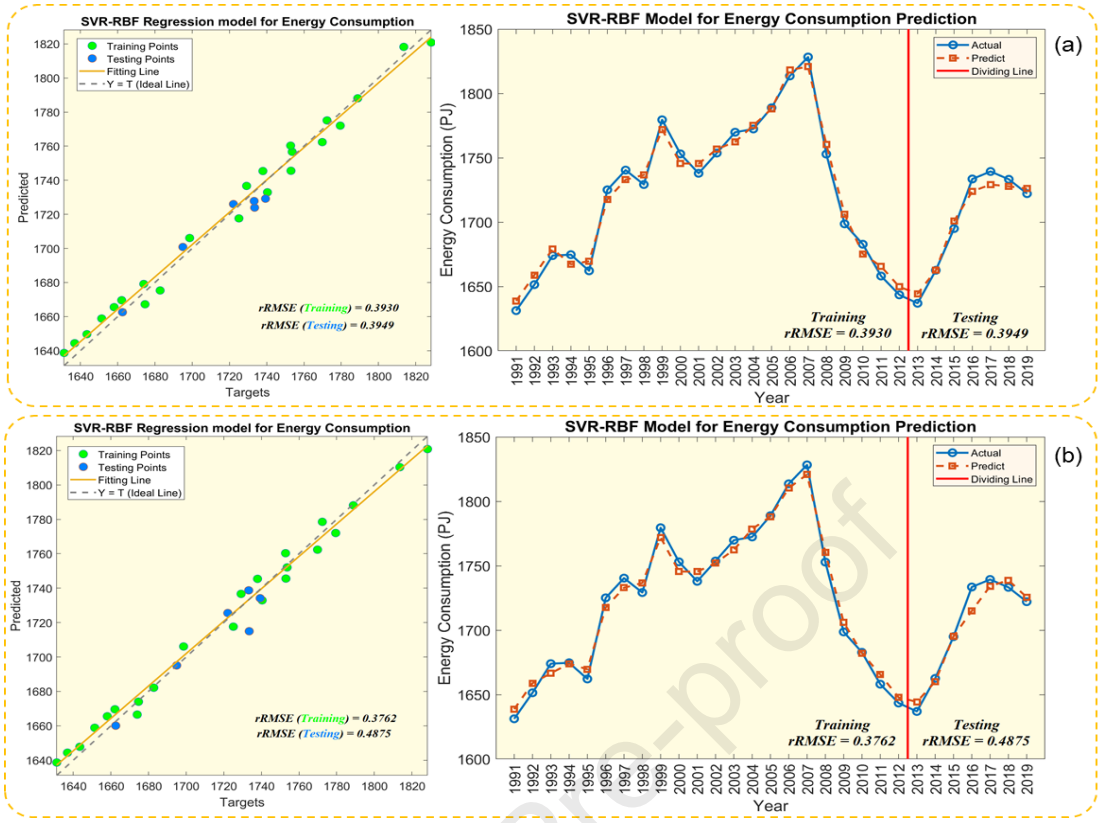


Figure 10. Performance of SVR-RBF in forecasting EC-Trans(t+1) based on (a) S1 and (b) S2 feature subsets.

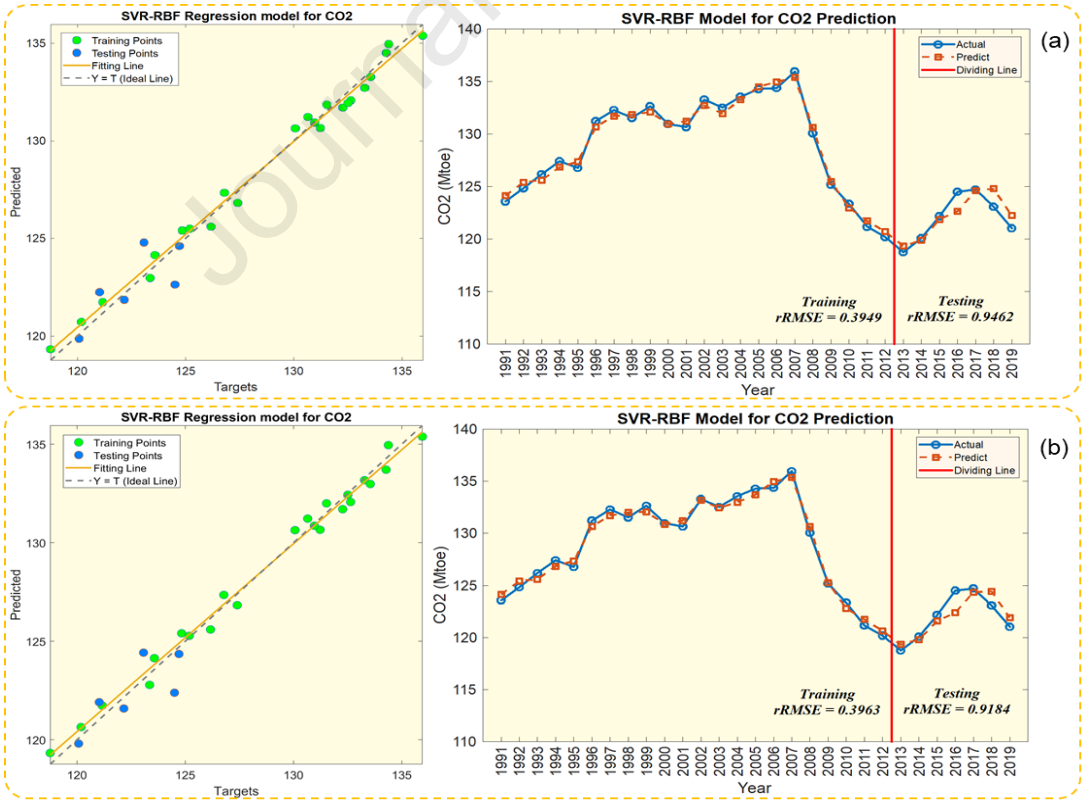


Figure 11. Performance of SVR-RBF in forecasting CO2(*t*+1) based on (a) S1 and (b) S2 feature subsets.

5.5 SHAP analysis

The SHAP method was utilised to enhance the interpretability of ML models and examine the impact of input features on model output. Figures 12 to 14 summarised the SHAP values of the final determined features (i.e., S2 feature set) and their quantified contribution towards EC-Trans(t+1) and CO2(*t*+1) using SVR-RBF and RF, respectively. The SHAP graphs in Figures 12-15 (a) display each input feature as a vertical bar on the x-axis, indicating its SHAP value and contribution to the models' output. SHAP values can be positive or negative, indicating whether a feature increased or decreased the model output. The SHAP values magnitude signifies the effect's intensity. The colour of each bar denotes the feature value in relation to the mean predicted value in the dataset. Blue signifies low values or negative effects, while red signifies high values or positive effects. The colour scheme aids in interpreting feature contributions and understanding the relationship between their values and the model's predictions. A simplified version of SHAP values were depicted in Figures 12-15 (b) which summarise the overall importance of each feature. In this study, the SHAP method was employed for interpreting SVR-RBF and RF in predicting EngCons and CO_2E, respectively. As shown in Figures 12 and 13, EC-trans played a significant role in EngCons while NVR, GDP and EC-All indicated a similar importance. In addition, it is observed that a higher value of EC-Trans and NVR led to a positive SHAP value and vice versa. Such association was not detected in GDP and EC-All. In terms of CO_2E, RCI contributed a significant higher proportion than any other features (i.e., CO_2, EC-All, GP and NEI). Similarly, association between a higher value of RCI, CO_2 and a positive SHAP value is obtained and vice versa. For forecasting EC-Trans(t+1) based on RF, RCI, EC-Trans ranked top 2 and were significantly more important than the other 3 features. The relations between positive SHAP value and the higher value of RCI, EC-Trans and NVR is also notice in the Figure 14. UR however suggested an opposite result where a higher value is associated with negative SHAP value. While for CO_2E prediction, RCI was also regarded as the most critical feature, followed by CO2(t) and R&D. Similar associations are seen between greater RCI, CO2(t) values and positive SHAP values, and vice versa. An opposite result is detected in R&D.

Journal Pre-proof

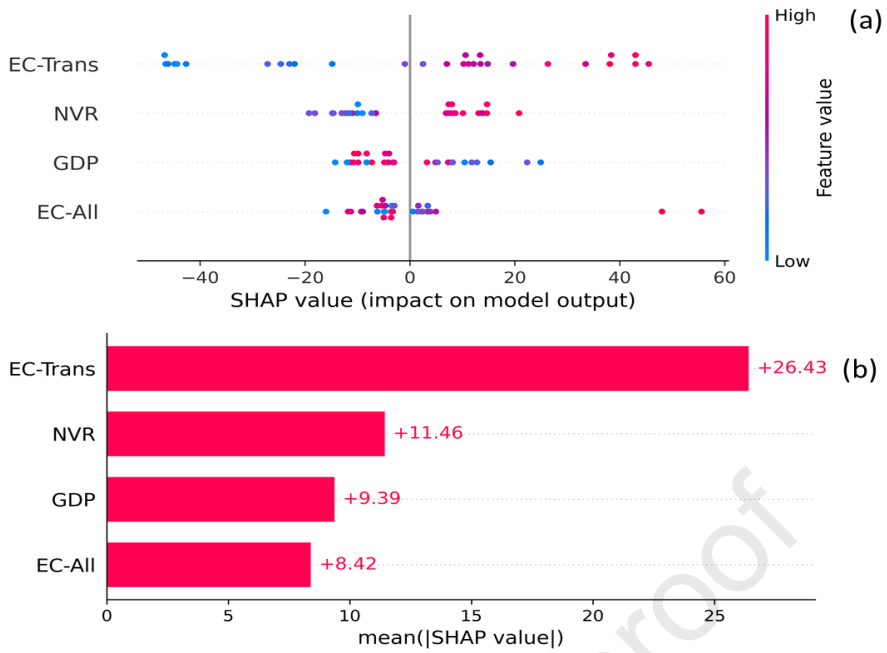


Fig. 12. The Shapley analysis for SVR-RBF for forecasting EC-Trans(t+1). (a) SHAP summary plot with S2 features, (b) The average contributions of the S2 features.

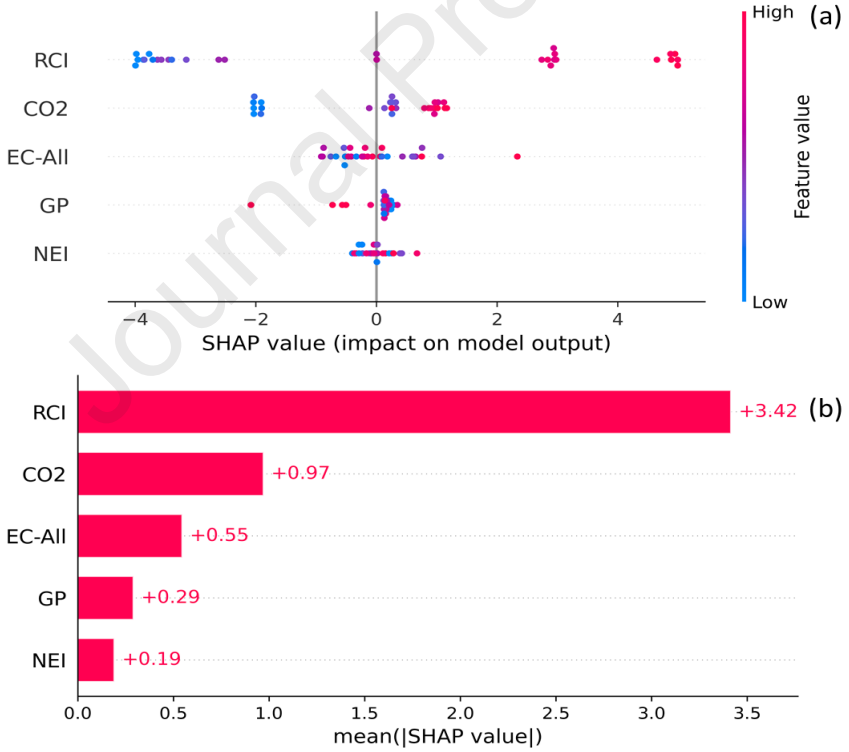


Fig. 13. The Shapley analysis for SVR-RBF for forecasting CO2(t+1). (a) SHAP summary plot with S2 features, (b) The average contributions of the S2 features.

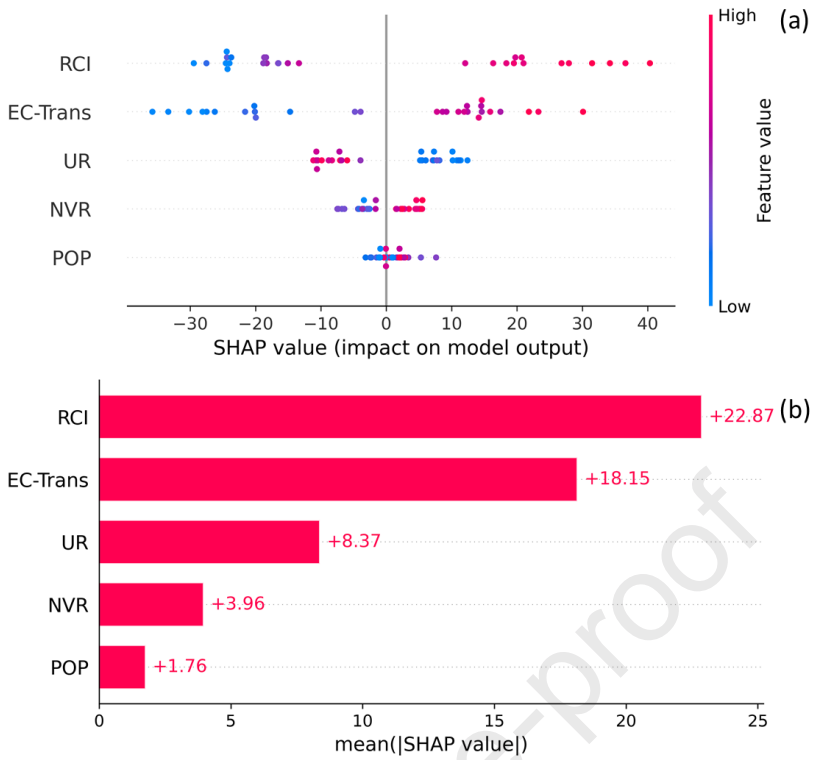


Figure 14. The Shapley analysis of RF for forecasting EC-Trans(t+1). (a) SHAP summary plot with S2 features, (b) The average contributions of the S2 features.

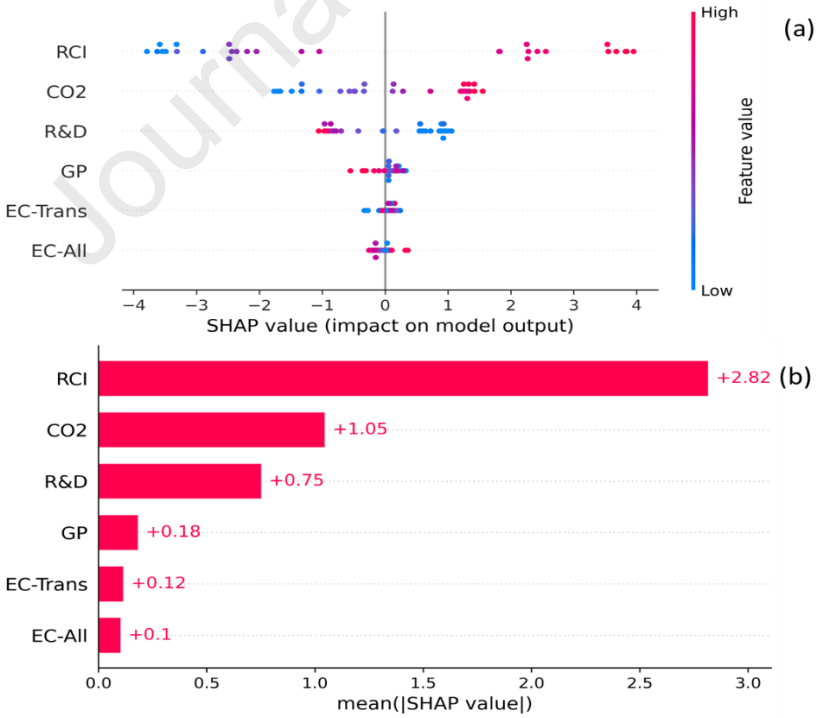


Fig. 15. The Shapley analysis of RF for forecasting CO2(t+1). (a) SHAP summary plot with S2 features, (b) The average contributions of the S2 features.

6. Noticeable discussions and limitations

A few noticeable points and limitations have been figured out in this study which are worth revealing for future similar studies. The dataset comprises a small size of 30 observations in the 1990-2019 period with 24 features, thereby such a small sample size leads to some challenges for MLs such as overfitting and the presence of random effects, which may negatively disturb the generalisation capability of ML models. To minimise the adverse impact caused by a small sample size, multiple FS stages within the forecasting framework was proposed to select the most related features and the feature space was therefore significantly reduced from initial 22 input features to less than 10 features. The risk of sparse matrix was then avoided.

Wrapper and embedded FS methods are both ML-based approaches. However, as previously mentioned, the issue of overfitting can hinder MLs from producing the most representative subset of features. In the context of filter methods, it is possible for two features to present a strong correlation based solely on numerical values, but this correlation may not essentially hold true. Therefore, in this paper, we proposed a voting mechanism to achieve a common agreement from three popular FS methods (i.e., mRMR, Boruta and RF). The preference of individual feature selection methods was mitigated to a greater extent.

It has been observed that most ML models exhibit a decrease in performance because of reducing the dimensionality of the feature set. This is comprehensible as the removed feature(s) may still possess a nuanced amount of valuable information but were excluded due to significant multicollinearity. The potential influence of the limited scope of the included data on this matter should not be underestimated. In comparison to high dimension datasets, such as gene analysis, low dimension dataset is less prone to containing redundant or irrelevant variables. However, this does not imply that the proposed FS methods were incapable of handling low-dimensional data, as suggested by SVR-RBF methods. In fact, a superior performance was achieved even with a reduced number of features. Also, it is important to note that despite a compromise in forecasting accuracy, the proposed framework retains the most related features for EngCons in transport sector, in which policymakers can benefit more from correct and accurate conclusion rather than accurate prediction but contain misleading information.

Lastly in all studies, GDP and population were selected as key driving variables for forecasting EC-Trans(t) and CO2(t), but in this study neither GDP nor population are found as influential variables in the UK's transportation sector, possibly because the trend of both EngCons and CO_2E are rapidly decreasing in the UK.

7. Conclusions and future research

A sustainable transport system necessitates a comprehension of the associations between transport EngCons and CO_2E, and their contributing factors, which also facilitates achieving promising performance of MLs in forecasting. To this end, this paper proposes an interpretable multi-stage forecasting framework to quantify EngCons and CO_2E in the UK's transport sector and identifying the

most relevant factors based on 22 initial input features from multisource including socioeconomic, transportation- and energy-related variables. Unlike recent published papers that solely focused on achieving the best prediction accuracy, the proposed framework also integrated interpretable ML methods to simultaneously maximise the forecasting accuracy and to determine the relationship between the forecasts and the influential variables using the SHAP method.

The contributions of this paper are as follows:

- To the best of our knowledge, this study is the first attempt, which employed a large list of input features and performed correlation and multicollinearity analyses to remove highly correlated features to provide an appropriate subset of features for interpretability of black-box ML models.
- This study introduces a novel voting scheme for feature selection (FS), which combines both filter and embedded paradigms, which has not been studied before in the EngCons context.
- This study is the second work in EngCons context (the first study is [30]) that applies the SHAP analysis to forecast the EngCons and CO2E to determine the influential variables.

The results indicate that the proposed multi-stage FS framework was able to improve the quality of data by removing potentially irrelevant and redundant features, in which average *rRMSE* and average *MAPE* of 1.024 and 0.915 for forecasting EC-Trans(t+1), and average *rRMSE* and average *MAPE* of 1.073 and 0.907 for forecasting CO2(t+1) with S1 feature subsets are achived. The selected best ML model varies depending on the feature subset examined. Overall, in both S1 and S2 feature subsets SVR-RBF and LSTM have the best and the weakest performance.

Shapley analysis for UK's transport EngCons and CO_2E forecasting indicates that road carbon intensity is the most significant factor associated with both EngCons and CO_2E. Unlike similar studies where population and GDP are key driving variables, Shapley analysis reveals that only GDP is selected as contributing variable for forecasting EC-Trans(t+1).

In 2020 the United Nations Economic Commission for Europe recommended that countries investigate the possibility of reporting quarterly GHG emissions data as part of climate change statistics [63]. There are solid methodologies for estimating GHG emissions on an annual basis and a few countries currently strived to develop statistical methodologies to compile quarterly time series emissions. Thus, for future researchers are encouraged to use quarterly GHG emissions data, rather than annual observations in which ML models avoid facing challenges with small sample size.

References

[1] Ahmed Z, Asghar MM, Malik MN, Nawaz K. Moving towards a sustainable environment: the dynamic linkage between natural resources, human capital, urbanization, economic growth, and ecological footprint in China. Resources Policy. 2020; 67:101677.

[2] Ağbulut Ü. Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. Sustainable Production and Consumption. 2022; 29:141-57.

[3] GreenhouseGas.Statistics. UK territorial greenhouse gas emissions national statistics. 2021.

[4] Energy.Stats. Energy Consumption in the UK (ECUK) 1970 to 2021. 2022.

[5] IEA. Global CO2 emissions in transport by mode in the sustainable development scenario, 2000-2070. 2022.

[6] Maaouane M, Zouggar S, Krajačić G, Zahboune H. Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. Energy. 2021; 225:120270.

[7] Öztürk OB, Başar E. Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. Ocean Engineering. 2022; 243:110209.

[8] Wang M, Wang Y, Chen L, Yang Y, Li X. Carbon emission of energy consumption of the electric vehicle development scenario. Environmental Science and Pollution Research. 2021; 28:42401-13.

[9] Garcia J, Teodoro F, Cerdeira R, Coelho L, Kumar P, Carvalho M. Developing a methodology to predict PM10 concentrations in urban areas using generalized linear models. Environmental technology. 2016; 37(18):2316-25.

[10] Adebayo TS, Awosusi AA, Kirikkaleli D, Akinsola GD, Mwamba MN. Can CO2 emissions and energy consumption determine the economic performance of South Korea? A time series analysis. Environmental Science and Pollution Research. 2021; 28(29):38969-84.

[11] Shahbaz M, Loganathan N, Sbia R, Afza T. The effect of urbanization, affluence and trade openness on energy consumption: A time series analysis in Malaysia. Renewable and Sustainable Energy Reviews. 2015; 47:683-93.

[12] Hu Y-C, Jiang P. Forecasting energy demand using neural-network-based grey residual modification models. Journal of the Operational Research Society. 2017; 68:556-65.

[13] Maaouane M, Chennaif M, Zouggar S, Krajačić G, Duić N, Zahboune H, ElMiad AK. Using neural network modelling for estimation and forecasting of transport sector energy demand in developing countries. Energy Conversion and Management. 2022; 258:115556.

[14] Ekonomou L. Greek long-term energy consumption prediction using artificial neural networks. Energy. 2010; 35(2):512-7.

[15] Jana RK, Ghosh I, Sanyal MK. A granular deep learning approach for predicting energy consumption. Applied Soft Computing. 2020; 89:106091.

[16] Pan Y, Fang W, Zhang W. Development of an energy consumption prediction model for battery electric vehicles in real-world driving: A combined approach of short-trip segment division and deep learning. Journal of Cleaner Production. 2023; 400:136742.

[17] Lei L, Chen W, Wu B, Chen C, Liu W. A building energy consumption prediction model based on rough set theory and deep learning algorithms. Energy and Buildings. 2021; 240:110886.

[18] Kaytez F. A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption. Energy. 2020; 197:117200.

[19] Wang X, Luo D, Zhao X, Sun Z. Estimates of energy consumption in China using a self-adaptive multi-verse optimizer-based support vector machine with rolling cross-validation. Energy. 2018; 152:539-48.

[20] Ullah I, Liu K, Yamamoto T, Zahid M, Jamal A. Electric vehicle energy consumption prediction using stacked generalization: An ensemble learning approach. International Journal of Green Energy. 2021; 18(9):896-909.

[21] Amiri SS, Mostafavi N, Lee ER, Hoque S. Machine learning approaches for predicting household transportation energy use. City and Environment Interactions. 2020; 7:100044.

[22] Al-Ghandoor A, Samhouri M, Al-Hinti I, Jaber J, Al-Rawashdeh M. Projection of future transport energy demand of Jordan using adaptive neuro-fuzzy technique. Energy. 2012; 38(1):128-35.

[23] Sahraei MA, Çodur MK. Prediction of transportation energy demand by novel hybrid meta-heuristic ANN. Energy. 2022; 249:123735.

[24] Piecyk MI, McKinnon AC. Forecasting the carbon footprint of road freight transport in 2020. International journal of production economics. 2010; 128(1):31-42.

[25] Logan KG, Nelson JD, Chapman JD, Milne J, Hastings A. Decarbonising UK transport: Implications for electricity generation, land use and policy. Transportation Research Interdisciplinary Perspectives. 2023; 17:100736.

[26] UK I. UK Transport Vision 2050: investing in the future of mobility. 2021.

[27] Mitrentsis G, Lens H. An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. Applied Energy. 2022; 309:118473.

[28] Amiri SS, Mueller M, Hoque S. Investigating the application of a commercial and residential energy consumption prediction model for urban Planning scenarios with Machine Learning and Shapley Additive explanation methods. Energy and Buildings. 2023; 287:112965.

[29] Pokharel S, Sah P, Ganta D. Improved prediction of total energy consumption and feature analysis in electric vehicles using machine learning and shapley additive explanations method. World Electric Vehicle Journal. 2021; 12(3):94.

[30] Aras S, Van MH. An interpretable forecasting framework for energy consumption and CO2 emissions. Applied Energy. 2022; 328:120163.

[31] Wang L, Xue X, Zhao Z, Wang Y, Zeng Z. Finding the de-carbonization potentials in the transport sector: application of scenario analysis with a hybrid prediction model. Environmental Science and Pollution Research. 2020; 27(17):21762-76.

[32] Jurado S, Nebot À, Mugica F, Avellana N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. Energy. 2015; 86:276-91.

[33] Feng C, Cui M, Hodge B-M, Zhang J. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. Applied Energy. 2017; 190:1245-57.

[34] Zhang L, Wen J. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. Energy and Buildings. 2019; 183:428-42.

[35] Moldovan D, Slowik A. Energy consumption prediction of appliances using machine learning and multi-objective binary grey wolf optimization for feature selection. Applied Soft Computing. 2021; 111:107745.

[36] Qiao Q, Yunusa-Kaltungo A, Edwards RE. Feature selection strategy for machine learning methods in building energy consumption prediction. Energy Reports. 2022; 8:13621-54.

[37] Lv S-X, Wang L. Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model. Energy. 2023; 263:126100.

[38] Javanmard ME, Tang Y, Wang Z, Tontiwachwuthikul P. Forecast energy demand, CO2 emissions and energy resource impacts for the transportation sector. Applied Energy. 2023; 338:120830.

[39] Wang L, Xue X, Zhao Z, Wang Y, Zeng Z. Finding the de-carbonization potentials in the transport sector: application of scenario analysis with a hybrid prediction model. Environmental Science and Pollution Research. 2020; 27:21762-76.

[40] Sahraei MA, Duman H, Çodur MY, Eyduran E. Prediction of transportation energy demand: multivariate adaptive regression splines. Energy. 2021; 224:120090.

[41] Li Z, Zhou B, Hensher DA. Forecasting automobile gasoline demand in Australia using machine learning-based regression. Energy. 2022; 239:122312.

[42] Li X, Ren A, Li Q. Exploring Patterns of Transportation-Related CO2 Emissions Using Machine Learning Methods. Sustainability. 2022; 14(8):4588.

[43] Korkmaz E. Energy demand estimation in Turkey according to modes of transportation: Bezier search differential evolution and black widow optimization algorithms-based model development and application. Neural Computing and Applications. 2023:1-22.

[44] Khajavi H, Rastgoo A. Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. Sustainable Cities and Society. 2023; 93:104503.

[45] Limanond T, Jomnonkwao S, Srikaew A. Projection of future transport energy demand of Thailand. Energy policy. 2011; 39(5):2754-63.

[46] Hill G, Heidrich O, Creutzig F, Blythe P. The role of electric vehicles in near-term mitigation pathways and achieving the UK's carbon budget. Applied Energy. 2019; 251:113111.

[47] Xie M, Wu L, Li B, Li Z. A novel hybrid multivariate nonlinear grey model for forecasting the traffic-related emissions. Applied Mathematical Modelling. 2020; 77:1242-54.

[48] Rasool Y, Zaidi SAH, Zafar MW. Determinants of carbon emissions in Pakistan's transport sector. Environmental Science and Pollution Research. 2019; 26(22):22907-21.

[49] Lu X, Ota K, Dong M, Yu C, Jin H. Predicting Transportation Carbon Emission with Urban Big Data. IEEE Transactions on Sustainable Computing. 2017; 2(4):333-44.

[50] Giannakis E, Serghides D, Dimitriou S, Zittis G. Land transport CO2 emissions and climate change: evidence from Cyprus. International Journal of Sustainable Energy. 2020; 39(7):634-47.

[51] Çodur MY, Ünal A. An estimation of transport energy demand in Turkey via artificial neural networks. Promet-Traffic&Transportation. 2019; 31(2):151-61.

[52] Hanchuan P, Fuhui L, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(8):1226-38.

[53] Rogers J, Gunn S. Identifying Feature Relevance Using a Random Forest. Springer Berlin Heidelberg; 2006. p. 173-84.

[54] Qiao Q, Yunusa-Kaltungo A. A hybrid agent-based machine learning method for human-centred energy consumption prediction. Energy and Buildings. 2023; 283:112797.

[55] Mansfiled E, Billy P. Detecting Multicollinearity. The American Statisticians, 36, 158-160. 1982.

[56] Thompson CG, Kim RS, Aloe AM, Becker BJ. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. Basic and Applied Social Psychology. 2017; 39(2):81-90.

[57] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowledge and Information Systems. 2008; 14(1):1-37.

[58] Rasmussen CE. Gaussian Processes in Machine Learning. Springer Berlin Heidelberg; 2004. p. 63-71.

[59] Palak M, Revati G, Hossain MA, Sheikh A. Deep learning models for smart building load profile prediction. Conference Deep learning models for smart building load profile prediction. IEEE, p. 1-6.

[60] Ardabili S, Abdolalizadeh L, Mako C, Torok B, Mosavi A. Systematic Review of Deep Learning and Machine Learning for Building Energy. Frontiers in Energy Research. 2022; 10.

[61] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001; 29(5):1189-232.

[62] Qiao Q, Yunusa-Kaltungo A, Edwards RE. Towards developing a systematic knowledge trend for building energy consumption prediction. Journal of Building Engineering. 2021; 35:101967.

[63] Saadatmand H, Akbarzadeh-T M-R. Set-based integer-coded fuzzy granular evolutionary algorithms for high-dimensional feature selection. Applied Soft Computing. 2023; 142:110240.

[64] Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv pre-print server. 2017.

[65] Liu X, Tang H, Ding Y, Yan D. Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings. Energy and Buildings. 2022; 273:112408.

Appendix A

Table A1. Descriptive statistics for the full features

Feature Abbreviation	Unit	Mean	Std	Min	Max
CO2	Mtoe	127.39	5.09	118.76	135.96
EC-Trans	PJ	1716.13	54.84	1631.21	1828.29
OP-Trans	PJ	1675.92	61.87	1577.97	1798.86
EC-All	PJ	5861.52	394.87	5144.19	6341.47
TES	PJ	8625.08	741.32	7145.14	9449.86
POP	Million	8.90	9.80	2.00	9.00
UPR	%	80.20	1.87	78.11	83.65
EI	MJ/pkm	144.04	31.99	94.13	190.43
GDP	USD	34816.21	10197.2	18389.02	50653.26
EC-Elec	PJ	1130.26	76.32	987.96	1255.23
RW-Trans	PJ	18.79	22.54	0.00	68.90
SEV	%	0.47	0.81	0.00	3.30
UR	%	6.54	1.82	3.80	10.40
GP	USD	1.41	0.46	0.80	2.20
R&D	USD	419.23	358.78	69.90	1290.15
NEI	PJ	971.17	1991.90	-2014.40	4026.40
RCI	gCO2/MJ	70.80	0.57	69.30	71.60
AP	Million	91.32	34.34	42.86	165.39
AF	Million tone	5702.52	847.88	3825.40	7618.10
RP	Million	1169.33	354.97	735	1744
NVR	Thousand	2714.65	402.78	1901.80	3295.96
TV	Million	31345.42	4606.01	24511.00	38682.70
AVM	1000 miles	9.65	0.59	8.90	10.53
TVM	Billion vehicle miles	295.88	26.28	241.00	338.60

Journal Pre-proof

Nomenclature

Symbol	Description
I	Mutual information
p	Joint probabilistic density
S	Data set
$maxD$	Max-Relevance
$minR$	Minimum-Redundancy
x, X	Input feature, input vector
V_{mf}	Whether select feature or not
V_T	The total vote threshold value
$\mathcal{D}$	Training dataset
$y, \bar{y}, \tilde{y}/p$	Actual, average actual and predicted target variable
$\mathcal{N}$	Gaussian distribution
σ ()	Sigmoid function
f_t	Forgetting gate
C ()	Cell state
i_t	Input gate
o_t	Output gate
h ()	Hidden state
W_f, W_i, W_o	Recurrent weighting metrics
b_f, bi, b_o	Bias vectors
$\beta_1, \beta_2, \beta_m$	Coefficients
ε	Random error
E	Expectation function
L	Loss function
a_m, p_m	Increment/step/boost of LSBoost
η	Learning rate of ANN
ϕ'	The derivative of the activation function
ω	Weights of ANN nodes
ϕ	SHAP value
ϕ_0	The mean value of the output variable

Journal Pre-proof

Table 4. Result of FS voting scheme for EC-Trans(t+1).

FS Method	CO2(t)	EC-Trans(t)	OP-Trans	EC-All	TES	R&D	UR	RCI	NVR	TV	Total count
mRMR	●	●	●	●	○	○	●	●	●	●	8
Boruta	●	●	●	●	●	○	●	●	●	○	8
RF	●	●	●	●	○	●	●	●	●	○	8
Score	3	3	3	3	1	1	3	3	3	1	10

Table 5. Result of FS voting scheme for CO2(t+1).

FS Method	CO2(t)	EC-Trans(t)	OP-Trans	EC-All	TES	RW-Trans	R&D	NEI	RCI	AVM	GP	POP	Total count
mRMR	●	●	●	●	●	●	●	●	●	○	○	●	10
Boruta	●	○	●	●	●	●	●	●	●	●	●	○	10
RF	●	●	●	●	●	●	●	●	●	●	○	○	10
Score	3	2	3	3	3	3	3	3	3	2	1	1	12

Table 6. Selected feature subsets with the ML models for forecasting EC-Trans(t+1).

Feature subset	ML models	Selected features												Number of features
		CO2(t)	OP-Trans	EC-All	TES	R&D	UR	RCI	NVR	TV	POP	GDP	EC-Trans(t)	
S1	LSBoost	○	●	●	○	●	○	○	●	●	○	●	●	7
	MPL	●	●	●	○	○	○	○	●	●	○	●	●	7
	SVR	○	○	○	○	○	○	●	●	●	●	○	●	5
	GPR	○	○	○	○	○	●	○	●	●	○	○	●	4
	LR	○	○	○	○	○	●	●	●	●	○	●	●	6
	RF	○	●	●	●	●	●	●	●	●	●	●	●	11
	LSTM	○	●	○	○	○	○	●	●	●	○	●	●	6
	Score	1	4	3	1	2	3	4	7	7	2	5	7	46
S2	LSBoost	N/A	N/A	●	N/A	●	●	○	●	N/A	N/A	○	●	5
	MPL			●		○	●	○	●			●	●	5
	SVR			●		○	○	○	●			●	●	4
	GPR			●		○	●	○	●			○	●	4
	LR			●		○	●	●	●			○	●	5
	RF			●		○	○	●	●			●	●	5
	LSTM			○		○	○	●	●			●	●	5
	Score			6		1	4	3	7			4	7	32

Note: N/A indicates the feature is excluded in S2

Journal Pre-proof

Table 7. Selected feature subsets with the ML models for forecasting $CO_2(t+1)$.

Feature subset	ML models	Selected features													Number of features
		EC-Trans(t)	OP-Trans	EC-All	TES	RW-Trans	R&D	NEI	RCI	AVM	GP	POP	GDP	CO2(t)	
S1	LSBoost	○	●	○	○	●	○	○	○	●	○	○	○	●	4
	MPL	●	○	○	○	○	●	○	●	●	●	○	●	●	7
	SVR	○	○	●	○	○	○	●	●	●	●	○	○	●	6
	GPR	○	○	○	○	○	●	○	○	●	●	○	○	●	4
	LR	●	○	●	○	○	●	○	●	●	●	○	●	●	8
	RF	○	●	●	○	●	○	○	●	●	●	○	○	●	7
	LSTM	●	●	○	●	○	●	●	●	●	●	●	●	●	11
	Score	3	3	3	1	2	4	2	5	7	6	1	3	7	47
S2	LSBoost	●	N/A	●	N/A	N/A	○	○	●	N/A	●	N/A	N/A	●	5
	MPL	●		●			○	●	●		●			●	6
	SVR	○		●			○	●	●		●			●	5
	GPR	●		○			●	●	●		●			●	6
	LR	●		●			●	○	●		●			●	6
	RF	●		●			●	○	●		●			●	6
	LSTM	●		●			○	●	●		●			●	6
	Score	6		6			3	4	7		7			7	40

Note: N/A indicates the feature is excluded in S2

Journal Pre-proof

Journal Pre-proof

Journal Pre-proof

Journal Pre-proof

Highlights:

- An interpretable multi-stage forecasting framework is proposed.
- The drivers of energy consumption and CO2 emissions in UK's transport sector are identified.
- A comprehensive comparison between different machine learning models is carried out.
- The impacts of feature selection methods on machine learning models are investigated.

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof