STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–19.
Published online in International Academic Press (www.IAPress.org)

Comparative Evaluation of AI Techniques for Vehicle Fuel Consumption and CO_2 Emission Prediction

Doha EL KALEF[1], Hanan SABBAR[1], Bouchaib CHERRADI[1,2,3],Hassan SILKAN[1]

[1]*LAROSERI Laboratory, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco*
[2]*EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco*
[3]*STIE Team, CRMEF Casablanca-Settat, provincial section of El Jadida, El Jadida, Morocco*

Abstract The transportation sector remains a major source of energy consumption and greenhouse gas emissions, making reliable vehicle-level estimates of fuel consumption and CO_2 emissions important for fleet management, regulatory assessment, and sustainable transport planning. However, predictive performance can be overstated when models are evaluated through random data splits or when target-related information leaks into the input features. This study therefore proposes a leakage-aware machine learning framework designed to assess generalization across time and data sources. Canadian vehicle records from 2015–2020 are used for model development, whereas records from 2021–2024 are reserved for temporal evaluation. External validity is examined using US EPA vehicle data. Linear regression, random forest, extra trees, support vector regression, LightGBM, XGBoost, CatBoost, and FT-Transformer are compared under a common protocol. All preprocessing steps are fitted only on training folds, and target-derived variables are excluded. PCA sensitivity, out-of-fold stacking, statistical testing, residual analysis, SHAP interpretation, and computational cost are also examined. The framework provides a reproducible basis for developing dependable vehicle-screening tools under temporal and cross-dataset distribution shifts.

Keywords Machine learning, fuel consumption prediction, CO_2 emission prediction, data leakage prevention, temporal validation, cross-dataset validation.

DOI: 10.19139/soic-2310-5070-3508

1. Introduction

The automotive industry is undergoing a transformation through artificial intelligence (AI) in manufacturing, engineering, maintenance, and services. AI methods support vehicle design, production planning, quality inspection, predictive maintenance, traffic analysis, and environmental assessment. Previous research has examined the general influence of AI on the automotive industry [1], its application to predictive maintenance and service optimization [2], and its contribution to automotive manufacturing processes [3]. These developments show that data-driven technologies can support more efficient and sustainable transportation systems.

This transition depends not only on technological progress but also on the institutional context of energy and climate action. International organizations such as the International Energy Agency contribute to global energy governance [4], while national climate strategies are increasingly developed in response to international climate assessments [5]. Within this context, transportation remains an important source of energy-related greenhouse gas emissions because of its continued dependence on fossil fuels [6]. Reducing its environmental impact therefore requires policies that combine cleaner technologies, improved vehicle efficiency, and effective emission-reduction measures [7]. Recent energy assessments emphasize the need to accelerate emission reductions from energy use [8].

At the vehicle level, fuel consumption and CO_2 emissions are influenced by several interacting characteristics, including engine displacement, number of cylinders, transmission configuration, fuel type, vehicle class,

*Correspondence to: Doha EL KALEF ,elkalef.d@ucd.ac.ma. LaROSERI Laboratory, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

and structural design. Reliable prediction of these quantities can support preliminary vehicle assessment, fleet management, regulatory analysis, vehicle selection, and environmental policymaking. Machine-learning models have demonstrated their ability to identify relationships between technical vehicle attributes and fuel consumption [9]. Similar approaches have also been investigated for predicting carbon emissions from light-duty vehicles [10]. Nevertheless, the practical value of these models depends not only on their reported accuracy but also on the experimental protocol used to obtain that accuracy.

Several methodological choices can lead to overly optimistic predictive performance. Random data splits may place vehicles from the same or adjacent model years in both the training and test sets, thereby measuring interpolation rather than generalization to future vehicles. Information leakage can also occur when preprocessing is fitted before data partitioning or when target-derived variables are retained as predictors. For example, city and highway consumption are components of combined fuel consumption, while measured fuel consumption is closely related to CO_2 emissions. PCA and stacking may also be reported as beneficial without model-specific sensitivity analysis or valid out-of-fold training.

This study addresses these concerns through a leakage-aware comparative framework for the separate prediction of combined fuel consumption and CO_2 emissions. Linear Regression, Random Forest, Extra Trees, Support Vector Regression, LightGBM, XGBoost, CatBoost, and FT-Transformer are evaluated under a common protocol. Target-derived and cross-target variables are excluded, and all preprocessing operations are fitted using training data only. Canadian vehicle records from 2015–2020 are used for model development, while the complete 2021–2024 period remains untouched until temporal testing. Cross-dataset generalization is assessed using independently sourced U.S. EPA vehicle records.

The contribution of this work is methodological rather than the introduction of another isolated prediction algorithm. First, it defines a more realistic prediction task based exclusively on independent vehicle specifications. Second, it uses chronological validation to assess future-year generalization instead of relying solely on random partitions. Third, PCA configurations are examined separately for each eligible model and target, while individual models, simple prediction averaging, and temporal out-of-fold stacking are compared without assuming that dimensionality reduction or ensembling will always improve performance. Fourth, model reliability is assessed through bootstrap confidence intervals, paired statistical tests, multi-seed stability, year-by-year evaluation, residual diagnostics, SHAP interpretation, training and inference costs, model size, and external validation. Integrating these elements within one reproducible framework provides a more demanding assessment of predictive reliability and supports preliminary vehicle evaluation by manufacturers, fleet managers, regulators, and policymakers.

The remainder of this article is organized as follows. Section 2 reviews the relevant literature and highlights the methodological limitations that remain to be addressed. Section 3 describes the datasets, models, and evaluation protocol used in this study. Section 4 presents and discusses the experimental results. Finally, Section 5 summarizes the main conclusions, outlines the study's limitations, and suggests directions for future research.

2. Literature Review

Several methodological developments are relevant to vehicle-level prediction. XGBoost constructs an additive ensemble of regularized trees and is effective for structured tabular data [11]. Principal component analysis provides a lower-dimensional representation of correlated variables [12], although its usefulness may depend on the learning algorithm. FT-Transformer applies feature tokenization and self-attention to numerical and categorical attributes [13], providing an attention-based alternative to conventional regression models.

Modern deep-learning models for tabular data have also been proposed as alternatives to conventional tree ensembles. TabNet uses sequential attention to select relevant features during decision steps [14], NODE represents differentiable oblivious decision-tree ensembles [15], and TabTransformer applies Transformer-based contextual embeddings to categorical variables [16]. These architectures are relevant because vehicle datasets contain heterogeneous numerical and categorical attributes. In the present study, FT-Transformer was retained as the representative attention-based tabular deep-learning baseline because it directly supports feature tokenization and categorical embeddings within the computational scope of the experiments.

Road-transport emissions have been examined using environmental assessment and benchmarking tools [17]. Deep-learning models combined with metaheuristic optimization have also been applied to CO_2-emission

prediction [18], while ensemble methods have demonstrated strong performance for fuel-economy estimation [19]. Regulatory vehicle data have enabled practical CO_2-prediction models [20], and explainable machine-learning methods have been used to identify influential vehicle characteristics [21]. Physically based models provide complementary insight by representing aerodynamic drag, rolling resistance, road gradient, and operating conditions [22].

Recent research has also investigated the environmental and public-health effects of vehicle pollutants [23], integrated frameworks for automotive operational performance [24], and embedded deep-learning systems for real-time vehicle analysis [25]. Although the latter works do not directly predict fuel consumption or CO_2 emissions, they illustrate the broader need for efficient, interpretable, and deployable automotive AI.

Many vehicle-prediction studies use data derived from the Natural Resources Canada Fuel Consumption Ratings database [26]. Their performance is generally reported using RMSE and MAE [27], together with the coefficient of determination R^2 [28]. These metrics describe complementary aspects of regression performance and should therefore be interpreted jointly, preferably with uncertainty estimates.

Al-Nefaie and Aldhyani applied deep learning to vehicle CO_2-emission prediction [29]. Alam *et al.* developed an explainable multilayer neural model [30], while Hien and Kor used linear and polynomial regression for fuel-consumption and CO_2 prediction [31]. These studies reported strong results, but their predictor configurations included fuel-consumption variables closely related to the outputs.

Related contributions published in *Statistics, Optimization & Information Computing* demonstrate the application of data-driven methods in adjacent sustainability domains. Pinto Vega *et al.* compared metaheuristic algorithms for minimizing energy losses and CO_2 emissions in AC microgrids [32]. Abdelsamiea *et al.* used gradient boosting to analyze relationships among health expenditure, economic indicators, and CO_2 emissions [33]. Boumais and Messaoudi compared CNN, LSTM, and Transformer architectures for short-term energy-load forecasting [34]. Although these studies address different applications, they confirm the value of optimization and machine learning for sustainability-related problems.

Further NRCan-based research evaluated CatBoost for CO_2 prediction [35], random forest and support vector regression for future-year fuel-consumption assessment [36], gradient boosting over an extended historical period [37], and a tuned random forest for Canadian light-duty vehicles [38]. Nevertheless, most previous studies relied on random partitions or retained city, highway, or combined fuel-consumption measurements among the predictors. Such variables are closely related to the targets and may simplify the prediction task.

Direct comparison of reported accuracy is therefore difficult. The studies differ in periods, sample sizes, predictors, targets, preprocessing, and validation. Models evaluated through random partitions benefit from training and test records drawn from similar year distributions, whereas a chronological split requires prediction for vehicles introduced after the development period. Similarly, estimating CO_2 emissions from measured fuel consumption is less demanding than estimating them exclusively from engine, transmission, fuel, manufacturer, and vehicle-class information. High R^2 values reported under these different conditions should therefore provide methodological context rather than a direct ranking of model quality. Robust assessment should also consider uncertainty, temporal stability, external transferability, and computational cost in practical deployment.

The literature therefore lacks a unified evaluation combining strict leakage control, temporal testing, model-specific PCA analysis, valid out-of-fold stacking, statistical uncertainty, interpretability, computational cost, and independent cross-dataset validation. The present study addresses these limitations by using independent vehicle specifications, temporally ordered model development, an untouched 2021–2024 Canadian test set, and external evaluation on U.S. EPA data.

3. Methodology and Experimental Framework

This section describes the methodological framework developed for predicting vehicle fuel consumption and CO_2 emissions. It covers data acquisition, leakage-aware preprocessing, chronological partitioning, model development, temporal hyperparameter selection, out-of-fold stacking, statistical evaluation, external validation, model interpretation, and computational-cost assessment. The complete workflow of the study is illustrated in Figure 1.

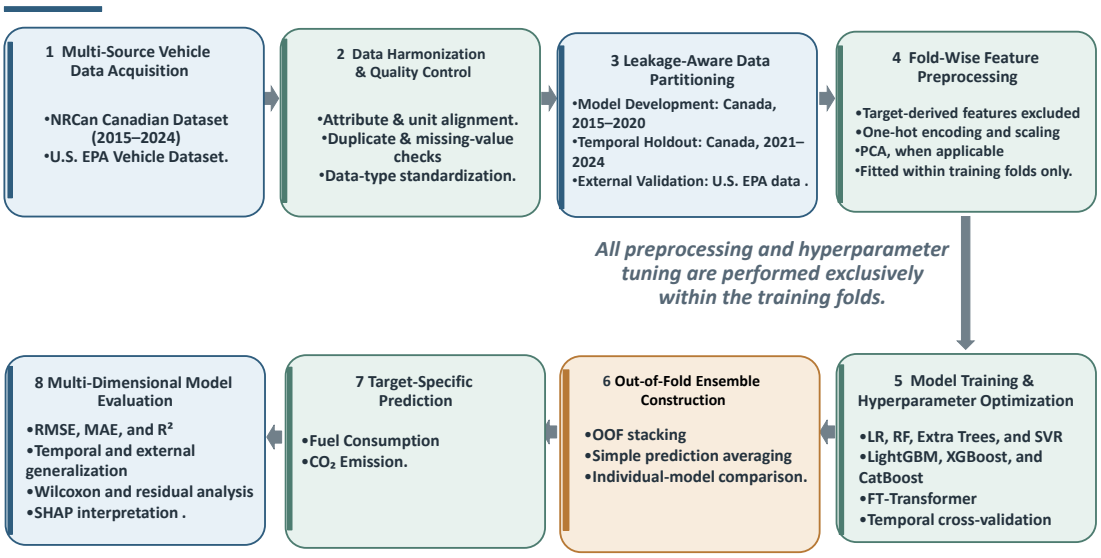


Figure 1. Leakage-aware temporal and cross-dataset workflow for vehicle fuel-consumption and CO_2-emission prediction.

3.1. Dataset and Prediction Tasks

The primary data were obtained from the official Natural Resources Canada Fuel Consumption Ratings database [26]. The dataset contains 10,060 light-duty vehicles from model years 2015–2024, described by 15 variables. These variables include model year, make, model, vehicle class, engine size, cylinder count, transmission, fuel type, city and highway fuel consumption, combined fuel consumption in L/100 km and mpg, CO_2 emissions, CO_2 rating, and smog rating.

Two continuous prediction tasks were considered separately: combined fuel consumption in L/100 km and CO_2 emissions in g/km. In both tasks, the objective was to estimate the target from vehicle specifications rather than from measurements derived from the target itself. Fuel consumption was therefore never used to predict CO_2 emissions, and CO_2 emissions were never used to predict fuel consumption. The predictors retained for these tasks and the measures taken to prevent information leakage are described in the following subsection.

3.2. Leakage-Aware Preprocessing and Temporal Partitioning

Before training the models, the data preparation procedure was reviewed with particular attention to information leakage. This step was necessary because some variables in the original database are either derived from the prediction targets or unavailable for part of the study period. Using such information could produce favorable test results without providing a realistic indication of future-year performance.

The initial data-quality audit showed that the variables retained for model development contained no missing observations. Two columns outside the final predictor set presented incomplete information: the CO_2 rating contained 1,128 missing values, while the smog rating contained 2,234. The CO_2 rating was removed because it is an ordinal representation derived from the measured CO_2-emission value. The smog rating was also excluded because it was not reported consistently throughout the study period and was entirely unavailable for the 2015 model year.

The eligibility of the remaining variables was considered separately for each prediction task. When combined fuel consumption was the target, city and highway consumption, combined fuel economy in mpg, CO_2

emissions, and both environmental ratings were removed. When CO_2 emissions were predicted, all measured fuel-consumption variables and environmental ratings were excluded. This procedure produced the same predictor set for both tasks: model year, make, model, vehicle class, engine size, cylinder count, transmission, and fuel type. Fuel consumption and CO_2 emissions were treated as separate targets, and neither was used as an input for predicting the other.

Before defining the final predictor set, the relationships among the numerical variables were examined to identify potential target-derived and cross-target information. Figure 2 presents the correlation matrix obtained before leakage control. City, highway, and combined fuel consumption are strongly associated because combined consumption is calculated from the two driving-cycle measurements. A strong association can also be observed between fuel consumption and CO_2 emissions. The decision to remove these variables was not based solely on a correlation threshold. They were excluded because they are functionally related to the prediction targets and would not necessarily be available when evaluating a new vehicle from its technical specifications.

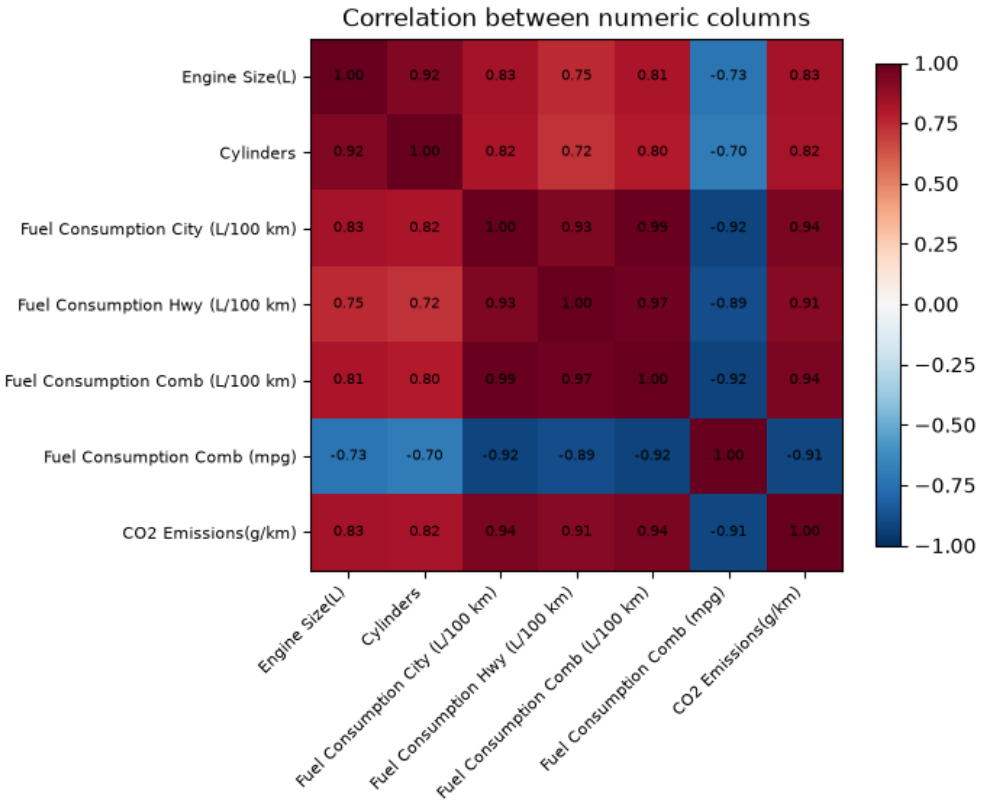


Figure 2. Correlation matrix of the numerical variables before leakage control. Target-derived and cross-target variables were excluded from the final predictor set.

The practical effect of excluding target-related information was also quantified. Figure 3 compares Linear Regression and XGBoost before and after leakage control under the same temporal test protocol. When target-derived and cross-target variables were retained, the resulting R^2 values were close to one for both prediction tasks. After these variables were removed, performance decreased to a more realistic level. This comparison confirms that the apparently near-perfect results obtained with the unrestricted predictor set were largely driven by information leakage rather than by genuine future-year generalization.

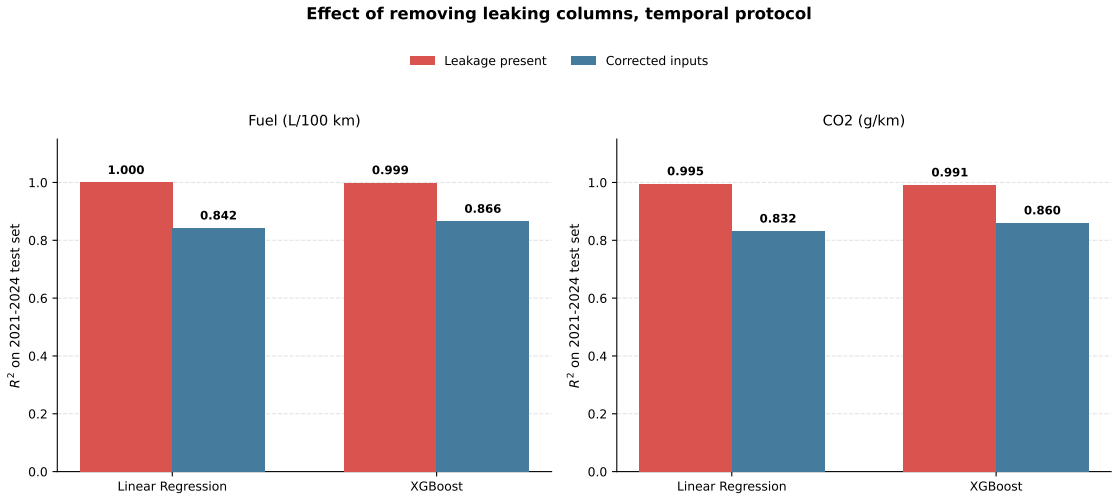


Figure 3. Effect of removing target-derived and cross-target predictors on the temporal-test performance of Linear Regression and XGBoost.

The data were then partitioned according to model year. Records from 2015–2020 ($n = 6{,}406$) were assigned to the development period, while records from 2021–2024 ($n = 3{,}654$) were reserved for temporal testing. The later period was not involved in preprocessing estimation, hyperparameter selection, PCA selection, early stopping, or meta-learner training. It was used only to report performance after model development had been completed.

Model selection within the development period followed an expanding-year validation scheme. The first fold used 2015–2016 for training and 2017 for validation. The training window was then expanded one year at a time, with 2018, 2019, and 2020 serving successively as validation years. Thus, every validation observation occurred later in time than the records used to train the corresponding model.

For linear regression, random forest, extra trees, SVR, LightGBM, and XGBoost, numerical variables were standardized and categorical variables were converted using one-hot encoding. Median imputation was retained inside the numerical pipeline as a precaution for other evaluation data, even though the selected Canadian predictors did not contain missing values. Unknown categorical levels were handled without refitting the encoder. CatBoost used the original categorical attributes directly, whereas the FT-Transformer represented each category through an integer code and a trainable embedding. Categories that were absent from the training data were assigned to a separate out-of-vocabulary level.

All preprocessing operations were carried out within the model pipelines. As a result, the imputation values, scaling parameters, category vocabularies, and PCA transformations were estimated again from the training portion of every fold. Validation and test records therefore had no influence on these parameters.

PCA was examined as a sensitivity experiment rather than imposed on every model. The six models based on numerical encoding were evaluated under four configurations: no PCA and PCA retaining 90%, 95%, or 99% of the variance. Hyperparameters were tuned independently for each configuration because the appropriate model settings may change after dimensionality reduction. The PCA configuration retained for each model and target was selected from the mean validation RMSE across the expanding-year folds. The temporal test results were not used for this decision. CatBoost and FT-Transformer were not included in the PCA comparison because CatBoost uses native categorical handling, whereas FT-Transformer represents numerical and categorical variables through feature-specific tokens and embeddings.

Multicollinearity was also examined after leakage control. For the retained numerical predictors, the variance inflation factors were 6.76 for engine size, 6.76 for cylinder count, and 1.00 for model year. The moderate collinearity between engine size and cylinder count is expected because both variables describe powertrain capacity. This collinearity does not create information leakage, but it can affect coefficient interpretation in linear models. For PCA-based configurations, the transformed components are orthogonal by construction. VIF was not interpreted

for the complete one-hot encoded categorical indicators because full dummy-variable groups are structurally dependent.

The main elements of the preprocessing and temporal partitioning procedure are summarized in Table 1.

Table 1. Summary of the leakage-aware preprocessing and temporal partitioning procedure.

Category	**Variables or period**	**Treatment**
Final predictors	Model year, make, model, vehicle class, engine size, cylinders, transmission, and fuel type	Used for both prediction tasks
Excluded information	City and highway consumption, combined mpg, alternate target, CO_2 rating, and smog rating	Removed before model fitting
Development period	Canadian vehicle records from 2015–2020 ($n = 6{,}406$)	Model development and selection
Internal validation	Four expanding-year folds ending in 2017, 2018, 2019, and 2020	Chronologically ordered validation
Preprocessing	Numerical and categorical predictors	Fold-wise scaling, imputation, and encoding
PCA experiment	No PCA, PCA-90%, PCA-95%, and PCA-99%	Selected separately by validation RMSE
Temporal test	Canadian vehicle records from 2021–2024 ($n = 3{,}654$)	Not used for model selection

3.3. Predictive Models

Eight individual regression models were evaluated: Linear Regression, Random Forest, Extra Trees, Support Vector Regression (SVR), LightGBM, XGBoost, CatBoost, and the FT-Transformer. Each algorithm was trained independently for fuel consumption and CO_2 emissions using the same temporal partition.

Machine-learning benchmarks. Linear Regression was included as an interpretable reference model. Random Forest and Extra Trees represented bagging-based ensembles, while SVR with a radial basis function kernel represented nonlinear kernel regression. LightGBM, XGBoost, and CatBoost were selected as gradient boosting methods for structured tabular data. CatBoost used native categorical processing, while the other machine-learning models used the preprocessing pipeline described in Section 3.2.

XGBoost constructs an additive sequence of regression trees by minimizing the following regularized objective [11]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \gamma T_t + \frac{1}{2}\lambda\|\mathbf{w}_t\|_2^2 + \alpha\|\mathbf{w}_t\|_1, \tag{1}$$

where $l(\cdot)$ is the squared-error loss, T_t is the number of leaves, $\mathbf{w}_t$ contains the leaf weights, and γ, λ, and α control model complexity and regularization.

FT-Transformer. The FT-Transformer was included to evaluate attention-based learning for tabular regression [13]. Each numerical feature was converted into an individual token through a learned linear transformation, whereas each categorical feature was represented using a trainable embedding. A learned classification token was prepended to the sequence, and its final representation was supplied to a regression head after processing by the Transformer encoder.

The candidate architectures used token dimensions of 64 or 96, two or three Transformer layers, four attention heads, and dropout rates of 0.1 or 0.2. AdamW optimization, a batch size of 256, a maximum of 60 epochs, and early stopping with a patience of eight epochs were used. The 2020 data were reserved as the temporal validation year for architecture and epoch selection. A newly initialized model was then trained on the complete 2015–2020 period for the selected number of epochs.

3.4. Temporal Out-of-Fold Stacking

The stacking ensemble combined XGBoost and the FT-Transformer as heterogeneous base regressors. To prevent information leakage, the meta-learners were trained exclusively on temporal out-of-fold predictions. For every validation year from 2017 to 2020, the base models were trained using only the preceding years and subsequently used to predict the next year.

The out-of-fold meta-level matrix was defined as

$$\mathbf{Z}_{\mathrm{OOF}} = \left[\hat{\mathbf{y}}_{\mathrm{XGB}}^{\mathrm{OOF}} \quad \hat{\mathbf{y}}_{\mathrm{FT}}^{\mathrm{OOF}}\right], \tag{2}$$

where each row contains predictions generated for an observation that was not used to fit either base model. Ridge regression and XGBoost were evaluated as alternative meta-learners, while the unweighted average of the base predictions was included as an ensemble baseline.

After constructing the meta-training matrix, both base models were refitted on the complete 2015–2020 period. Their 2021–2024 predictions were then supplied to the fitted meta-learners. An intentionally contaminated variant, trained on in-sample predictions, was retained only as a leakage diagnostic and was excluded from final model selection.

3.5. Model Selection and Evaluation

For the conventional machine-learning models, ten randomly sampled hyperparameter configurations were evaluated using the four expanding annual folds. The configuration minimizing mean temporal validation RMSE was selected. PCA configuration and model hyperparameters were evaluated within the same training-period protocol, and every preprocessing operation was refitted within each fold. The 2021–2024 temporal test set was not used for hyperparameter selection.

The hyperparameter search spaces were defined before model evaluation and were kept identical for both prediction targets. For Random Forest and Extra Trees, the search covered the number of trees, maximum depth, minimum leaf size, and feature subsampling. For SVR, the penalty parameter, kernel scale, and insensitive-loss margin were varied. For LightGBM and XGBoost, the search included the number of estimators, tree depth, learning rate, subsampling, and regularization-related parameters. CatBoost was tuned over the number of iterations, depth, learning rate, and L2 leaf regularization. Linear Regression was used without hyperparameter tuning. The FT-Transformer architecture was selected from a small grid over token dimension, number of Transformer layers, and dropout. Table 2 summarizes the search spaces used in the experiments.

Table 2. Hyperparameter search spaces used for model selection.

Model	**Candidate values**
Linear Regression	No tuned hyperparameters
Random Forest	`n_estimators` $\in \{200, 400\}$; `max_depth` $\in \{\text{None}, 16, 24\}$; `min_samples_leaf` $\in \{1, 2, 4\}$; `max_features` $\in \{\sqrt{\cdot}, 0.5, 1.0\}$
Extra Trees	`n_estimators` $\in \{200, 400\}$; `max_depth` $\in \{\text{None}, 16, 24\}$; `min_samples_leaf` $\in \{1, 2, 4\}$; `max_features` $\in \{\sqrt{\cdot}, 0.5, 1.0\}$
SVR	$C \in \{1, 10, 30\}$; $\gamma \in \{\text{scale}, 0.01, 0.03\}$; $\epsilon \in \{0.05, 0.1, 0.2\}$
LightGBM	`num_leaves` $\in \{31, 63, 127\}$; `max_depth` $\in \{-1, 8, 12\}$; `learning_rate` $\in \{0.03, 0.05, 0.1\}$; `n_estimators` $\in \{200, 400, 600\}$; `min_child_samples` $\in \{10, 20, 40\}$; `subsample` $\in \{0.8, 1.0\}$
XGBoost	`n_estimators` $\in \{200, 400, 600\}$; `max_depth` $\in \{4, 6, 8\}$; `learning_rate` $\in \{0.03, 0.05, 0.1\}$; `subsample` $\in \{0.8, 1.0\}$; `colsample_bytree` $\in \{0.7, 0.85, 1.0\}$; $\alpha \in \{0, 0.5\}$; $\lambda \in \{0.5, 1.5, 3.0\}$
CatBoost	`iterations` $\in \{300, 600\}$; `depth` $\in \{6, 8\}$; `learning_rate` $\in \{0.03, 0.05, 0.1\}$; `l2_leaf_reg` $\in \{1, 3, 9\}$
FT-Transformer	Token dimension $\in \{64, 96\}$; number of layers $\in \{2, 3\}$; dropout $\in \{0.1, 0.2\}$; AdamW optimizer; batch size $= 256$; maximum epochs $= 60$; early stopping patience $= 8$

Performance was measured using root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5}$$

Here, y_i and $\hat{y}_i$ are the measured and predicted values, respectively, and $\bar{y}$ is the mean measured value. RMSE and MAE were reported in L/100 km for fuel consumption and g/km for CO_2 emissions.

Paired Wilcoxon signed-rank tests were applied to the absolute errors of predefined model pairs. The resulting p-values were adjusted separately for each target using the Holm procedure, with a significance threshold of 0.05. Confidence intervals for the performance metrics and paired RMSE differences were estimated using 2,000 paired bootstrap resamples.

Model stability was evaluated over five seeds: 42, 101, 202, 303, and 404. The hyperparameter configurations were held constant so that the experiment measured sensitivity to stochastic initialization rather than variability caused by repeated model selection.

The computational environment and reproducibility settings used in the experiments are summarized in Table 3.

Table 3. Computational environment and reproducibility settings.

Component	**Specification**
Programming environment	Python 3.12.12
Data processing	NumPy 1.26.4; Pandas 3.0.3
Machine-learning libraries	Scikit-learn 1.9.0; XGBoost 3.3.0; LightGBM 4.7.0; CatBoost 1.2.10
Deep-learning environment	PyTorch 2.13.0; FT-Transformer executed on CPU
Statistical and interpretation tools	SciPy 1.17.1; Statsmodels 0.14.6; SHAP 0.49.1
Processor	AMD64 Family 25 Model 33 processor; 6 physical and 12 logical cores
System memory	15.93 GB RAM
Graphics processor	NVIDIA GeForce GTX 1660 SUPER
Model execution device	XGBoost: GPU; Linear Regression, Random Forest, Extra Trees, SVR, LightGBM, CatBoost, and FT-Transformer: CPU; stacking: mixed GPU/CPU execution
Primary random seed	42
Stability-analysis seeds	42, 101, 202, 303, and 404
Inference-time protocol	One warm-up pass followed by seven repetitions; median and interquartile range reported
Reproducibility resources	Exact dependencies, execution order, and source code provided in the public repository

3.6. External Validation, Interpretation, and Efficiency

External validation was performed using the U.S. Environmental Protection Agency vehicle database. Models were trained using Canadian vehicles from 2015–2020 and evaluated under two conditions. The matched-year experiment used EPA vehicles from 2015–2020 to assess geographic domain shift, whereas the second experiment used EPA vehicles from 2021–2024 to assess combined geographic and temporal shifts.

Only attributes available in both databases were retained. Vehicle model names were removed because of inconsistent naming conventions. Transmission, fuel type, and vehicle-class labels were harmonized, and EPA fuel economy was converted from mpg to L/100 km using

$$\text{Fuel Consumption} = \frac{235.215}{\text{mpg}}. \tag{6}$$

EPA CO_2 emissions were converted from g/mile to g/km by dividing by 1.60934. Models were refitted on the common Canadian feature subset before external evaluation.

Model interpretation was conducted using SHAP values obtained from the strongest standalone XGBoost models, which were selected without PCA for both targets. One-hot contributions were aggregated back to their original attributes so that importance was reported for make, vehicle class, transmission, fuel type, and the numerical characteristics rather than individual encoded levels. Residuals were examined overall and by test year, vehicle class, and fuel type.

Computational efficiency was assessed using training time, inference latency, incremental memory consumption, peak memory, and serialized model size. Each model was measured in a separate process. Inference timing included one warm-up pass followed by seven repetitions, from which the median and interquartile range were calculated.

4. Results and Discussion

4.1. Temporal Performance, PCA Sensitivity, and Stability

The eight individual regression models were first evaluated on the untouched 2021–2024 Canadian temporal test set. All models used the same leakage-controlled predictors and chronological partition. RMSE and MAE are reported in the physical units of each target, whereas R^2 measures the proportion of explained variance. The complete temporal-test results are reported in Table 4.

Table 4. Performance of the individual models on the 2021–2024 temporal test set.

	Fuel consumption			**CO_2 emissions**		
Model	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
Linear Regression	1.1337	0.7918	0.8429	26.3046	18.7503	0.8323
Random Forest	0.9509	0.5655	0.8894	22.4109	13.3308	0.8783
Extra Trees	0.9515	0.5392	0.8893	23.2863	14.4217	0.8686
SVR	0.9999	0.6583	0.8778	31.8625	21.6879	0.7540
LightGBM	1.0345	0.6903	0.8692	23.3172	15.3353	0.8682
XGBoost	0.9032	0.5810	0.9003	21.0965	13.8247	0.8921
CatBoost	0.9551	0.6228	0.8885	21.9977	14.6409	0.8827
FT-Transformer	0.9822	0.6562	0.8821	21.5979	14.3058	0.8870

For both prediction tasks, XGBoost achieved the strongest individual performance according to RMSE and R^2. For fuel consumption, it obtained an RMSE of 0.9032 L/100 km, an MAE of 0.5810 L/100 km, and an R^2 of 0.9003. For CO_2 emissions, the corresponding values were 21.0965 g/km, 13.8247 g/km, and 0.8921, respectively. Extra Trees nevertheless produced the lowest fuel-consumption MAE, while Random Forest obtained the lowest CO_2 MAE, showing that model ranking also depends on the selected error criterion.

To complement these point estimates with an assessment of sampling uncertainty, Figure 4 compares the temporal-test RMSE values and their 95% bootstrap confidence intervals.

The bootstrap point estimates confirm that XGBoost obtained the lowest individual-model RMSE for both fuel consumption and CO_2 emissions. However, small differences between models should not be considered meaningful when their bootstrap intervals overlap. The results also demonstrate that model performance depends on the prediction target and selected evaluation metric.

The effect of dimensionality reduction was subsequently evaluated using four preprocessing configurations: no PCA and PCA retaining 90%, 95%, or 99% of the variance estimated from each training fold. The comparison was performed for Linear Regression, Random Forest, Extra Trees, SVR, LightGBM, and XGBoost. CatBoost and FT-Transformer were excluded because they operate on native categorical representations and feature-specific tokens, respectively.

The PCA sensitivity results are presented in Figure 5.

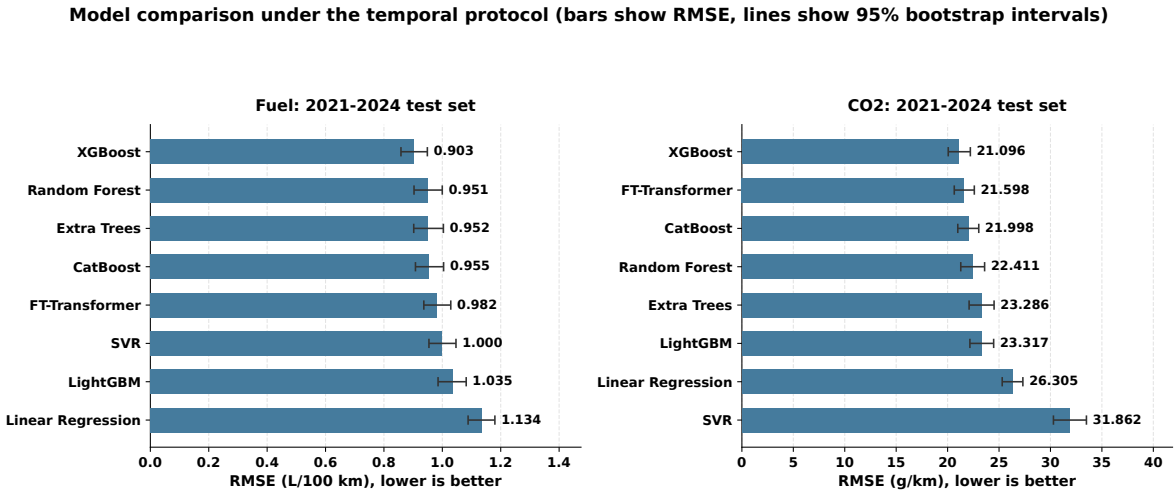


Figure 4. Temporal-test RMSE and 95% bootstrap confidence intervals for the eight individual regression models.

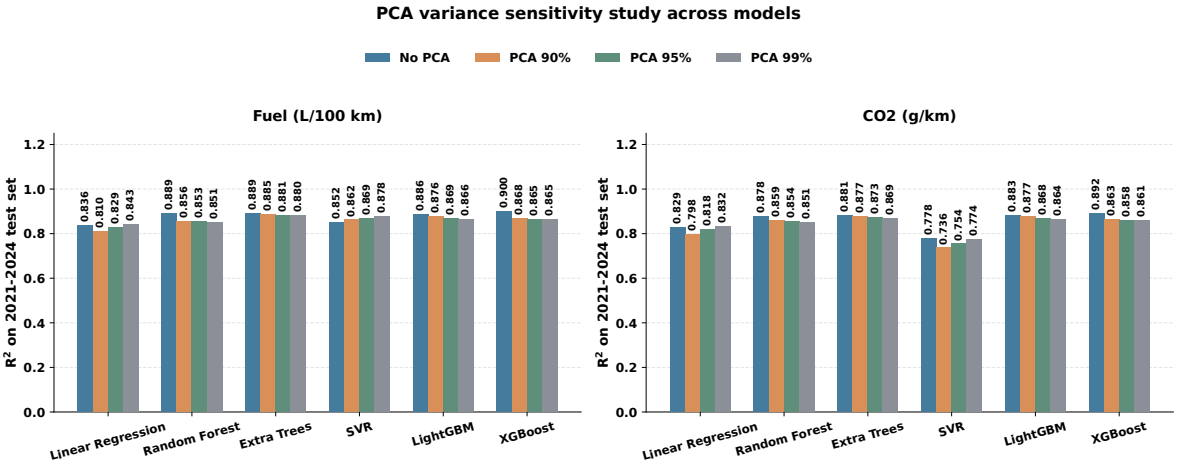


Figure 5. Sensitivity of the eligible regression models to the PCA configuration for fuel consumption and CO_2 emissions.

PCA did not provide a systematic improvement across all algorithms. Temporal validation selected PCA-99% for Linear Regression for both targets and for SVR in the fuel-consumption task. PCA-95% was selected for LightGBM for both targets and for SVR in the CO_2 task. Random Forest and XGBoost were selected without PCA for both targets, whereas Extra Trees was selected without PCA for fuel consumption and with PCA-99% for CO_2 emissions. For XGBoost, the no-PCA configuration exceeded PCA-95% by 0.0340 in R^2 for fuel consumption and by 0.0308 for CO_2 emissions.

These findings show that PCA should not be imposed uniformly. Although dimensionality reduction may remove redundant encoded dimensions and benefit linear or distance-based algorithms, it can also obscure informative thresholds and interactions used by tree ensembles. The final PCA configuration was therefore selected separately for each model and target using only temporal validation performance within the 2015–2020 training period.

To determine whether the principal model rankings were sensitive to algorithmic randomness, XGBoost, CatBoost, and FT-Transformer were retrained using five random seeds while their selected hyperparameters were held constant. The resulting mean and standard deviation of RMSE and R^2 are reported in Table 5.

Table 5. Multi-seed stability of the principal stochastic models on the 2021–2024 temporal test set.

Target	Model	RMSE mean ± SD	R^2 mean ± SD
Fuel	XGBoost	0.8994 ± 0.0045	0.9011 ± 0.0010
Fuel	CatBoost	0.9576 ± 0.0084	0.8879 ± 0.0020
Fuel	FT-Transformer	1.0571 ± 0.0499	0.8631 ± 0.0129
CO_2	XGBoost	21.1046 ± 0.1242	0.8921 ± 0.0013
CO_2	CatBoost	22.0220 ± 0.2642	0.8825 ± 0.0028
CO_2	FT-Transformer	23.1103 ± 1.5674	0.8700 ± 0.0178

XGBoost exhibited the lowest variability and retained the best mean performance for both targets across the five seeds. FT-Transformer showed the largest variation, particularly for CO_2 emissions, indicating moderate sensitivity to initialization and training stochasticity. The experiment therefore confirms the robustness of XGBoost's leading position, although the ordering of the lower-ranked models was less stable.

4.2. Stacking Ablation and Statistical Validation

The contribution of stacking was examined by comparing XGBoost and FT-Transformer individually, their unweighted prediction average, Ridge and XGBoost OOF meta-learners, and an intentionally contaminated reference. For the valid configurations, the meta-learners were fitted exclusively using temporally ordered out-of-fold predictions. The contaminated configuration was retained only as a leakage diagnostic and was excluded from final model selection. The numerical results of this ablation experiment are reported in Table 6.

Table 6. Stacking ablation results on the 2021–2024 temporal test set.

	Fuel consumption			CO_2 emissions		
Configuration	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**	R^2
XGBoost constituent	0.9032	0.5810	0.9003	21.0965	13.8247	0.8921
FT-Transformer constituent	0.9822	0.6562	0.8821	21.5979	14.3058	0.8870
Simple average	0.8992	0.5846	0.9012	20.2417	12.9901	0.9007
Ridge OOF stacking	0.8850	0.5722	0.9042	20.2415	13.1524	0.9007
XGBoost OOF stacking	0.9118	0.5884	0.8984	20.5992	13.1845	0.8972
Contaminated reference	0.8956	0.5802	0.9019	21.0979	13.3257	0.8921

Ridge OOF stacking produced the lowest valid RMSE for both targets. It achieved an RMSE of 0.8850 L/100 km and an R^2 of 0.9042 for fuel consumption, and an RMSE of 20.2415 g/km and an R^2 of 0.9007 for CO_2 emissions. Relative to the strongest constituent model, the RMSE decreased by approximately 2.0% for fuel consumption and 4.1% for CO_2 emissions.

The comparison shows that Ridge OOF stacking improved upon both constituent models for the two targets and also outperformed the simple average for fuel consumption. For CO_2, however, Ridge stacking and simple averaging produced nearly identical RMSE values, and the simple average obtained a slightly lower MAE. Thus, the benefit of stacking over a simple ensemble was target dependent. The contaminated configuration cannot support a generalization claim because its meta-training predictions were generated in sample.

To determine whether the observed performance differences were statistically supported, paired Wilcoxon signed-rank tests were applied to the absolute errors of predefined model pairs. Holm correction was performed separately for each target, while paired bootstrap resampling was used to estimate RMSE differences and their 95% confidence intervals. The resulting comparisons are reported in Table 7.

Table 7. Paired statistical comparison of the principal models. $\Delta\mathrm{RMSE} = \mathrm{RMSE}_A - \mathrm{RMSE}_B$; therefore, negative values favor Model A.

Target	Comparison	ΔRMSE [95% CI]	p_{Holm}
Fuel	CatBoost vs. Ridge OOF	$0.07012\,[0.05225,\ 0.08856]$	1.168×10^{-13}
Fuel	XGBoost vs. FT-Transformer	$-0.07900\,[-0.10940,\ -0.05134]$	5.789×10^{-21}
Fuel	XGBoost vs. Ridge OOF	$0.01816\,[0.01175,\ 0.02447]$	1.660×10^{-5}
Fuel	Ridge OOF vs. simple average	$-0.01417\,[-0.02375,\ -0.00462]$	1.660×10^{-5}
CO_2	CatBoost vs. Ridge OOF	$1.75620\,[1.43071,\ 2.09365]$	2.314×10^{-24}
CO_2	XGBoost vs. FT-Transformer	$-0.50137\,[-1.23111,\ 0.16723]$	0.6639
CO_2	XGBoost vs. Ridge OOF	$0.85498\,[0.62527,\ 1.05057]$	2.832×10^{-27}
CO_2	Ridge OOF vs. simple average	$-0.00016\,[-0.12431,\ 0.11466]$	5.276×10^{-7}

For fuel consumption, Ridge OOF stacking significantly outperformed XGBoost after Holm correction ($p_{\mathrm{Holm}} = 1.660 \times 10^{-5}$), and the confidence interval for the RMSE difference excluded zero. The same conclusion was obtained for CO_2 emissions ($p_{\mathrm{Holm}} = 2.832 \times 10^{-27}$).

For CO_2, the comparison between XGBoost and FT-Transformer was not statistically significant, and its RMSE-difference interval included zero. Although the Wilcoxon comparison between Ridge stacking and simple averaging was significant, their RMSE-difference interval $[-0.12431,\ 0.11466]$ included zero. This apparent difference arises because the Wilcoxon test evaluates paired absolute errors, whereas the bootstrap interval concerns the aggregate RMSE difference. The evidence therefore does not support a meaningful difference in CO_2 RMSE between these two ensemble strategies.

4.3. Temporal and Cross-Dataset Generalization

Performance was examined separately for each year of the Canadian test period to determine whether the aggregate metrics concealed temporal degradation. The annual performance of XGBoost, the strongest standalone model for both targets, is presented in Figure 6.

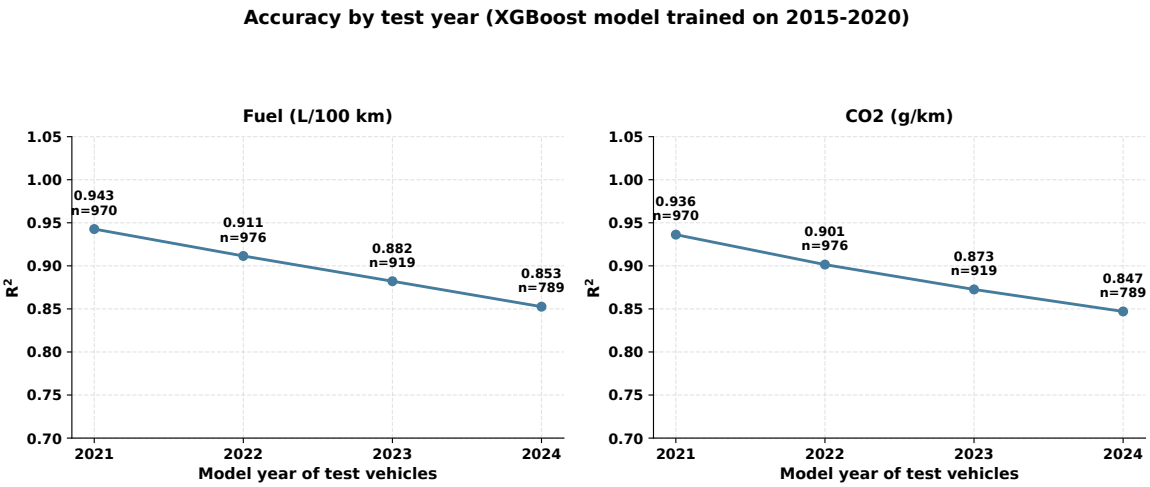


Figure 6. Year-by-year performance of the strongest standalone XGBoost models on the 2021–2024 Canadian temporal test period.

XGBoost achieved its strongest yearly performance in 2021, with $R^2 = 0.9427$ for fuel consumption and $R^2 = 0.9362$ for CO_2 emissions. Performance declined progressively, reaching $R^2 = 0.8526$ and $R^2 = 0.8470$, respectively, in 2024. Over the same period, fuel RMSE increased from 0.6907 to 1.0838 L/100 km, while CO_2 RMSE increased from 16.4111 to 25.1990 g/km. This pattern indicates measurable temporal degradation as the test year moves further away from the training period.

External generalization was subsequently assessed using independently sourced U.S. EPA vehicle records. The matched-year experiment evaluated geographic domain shift using EPA records from 2015–2020, whereas the second experiment evaluated the combined geographic and temporal shift using EPA records from 2021–2024. No EPA observations were used for model fitting, preprocessing estimation, hyperparameter selection, or meta-learner training. The external-validation results are reported in Table 8.

Table 8. Cross-dataset validation results obtained using the U.S. EPA vehicle data.

Experiment	**Target**	**Top model**	**n**	**RMSE**	**MAE**	**R^2**
Matched-year shift	Fuel	LightGBM	7,440	0.7826	0.4848	0.9055
Matched-year shift	CO_2	LightGBM	7,440	17.1203	10.4787	0.9162
Domain + future shift	Fuel	Ridge OOF stacking	4,484	1.0914	0.7601	0.8410
Domain + future shift	CO_2	Simple average	4,484	25.3287	17.4836	0.8435

Under the matched-year experiment, LightGBM achieved the lowest external RMSE for both fuel consumption (0.7826 L/100 km) and CO_2 emissions (17.1203 g/km). Under the combined geographic and future-year shift, Ridge OOF stacking obtained the lowest fuel-consumption RMSE of 1.0914 L/100 km, whereas the simple average obtained the lowest CO_2 RMSE of 25.3287 g/km.

For a visual comparison of the two external-validation scenarios, Figure 7 presents the performance obtained under matched-year geographic shift and combined geographic and future-year shift.

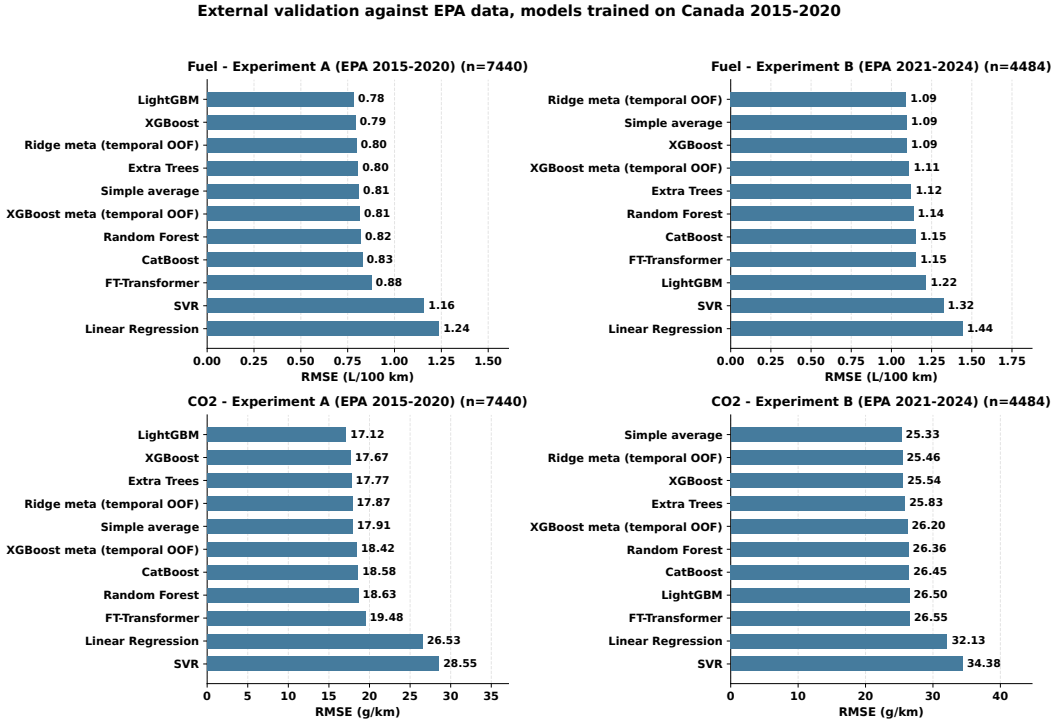


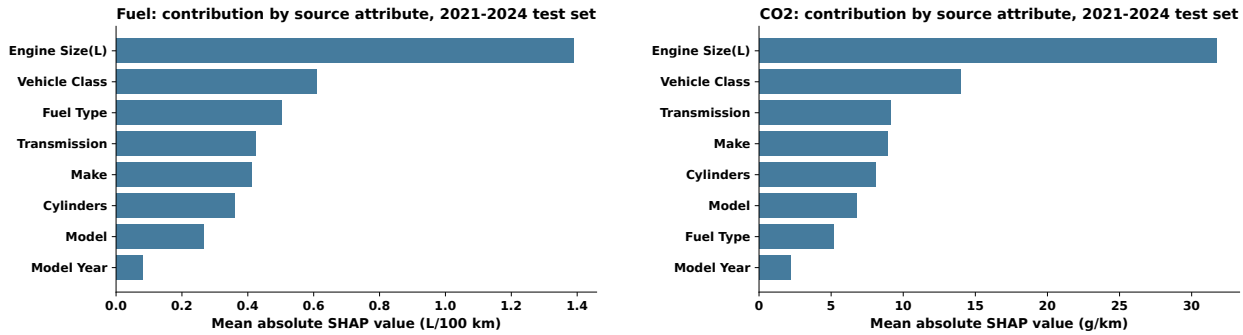
Figure 7. External validation on the U.S. EPA dataset under matched-year domain shift and combined domain and future-year shift.

Using XGBoost as a common reference across datasets, the matched-year EPA RMSE was 12.4% lower for fuel consumption and 16.3% lower for CO_2 than on the Canadian temporal test. In contrast, the EPA 2021–2024 RMSE was 21.2% higher for fuel and 21.1% higher for CO_2. The combined geographic and temporal shift therefore reduced transportability, although differences in target distributions and regulatory measurement procedures should also be considered.

These external results should not be interpreted as a direct comparison between Canadian and U.S. regulatory testing systems. Differences in vehicle specifications, category distributions, measurement protocols, and previously unseen categorical levels may affect transfer performance. Nevertheless, this experiment provides a more demanding assessment of model generalization than validation restricted to a single national dataset.

4.4. Interpretability, Computational Cost, and Comparison with Existing Studies

Feature attribution was computed for the strongest standalone XGBoost model for each target. The Ridge stacking model was not interpreted directly at the original-feature level because its meta-learner operates on base-model predictions. Contributions associated with one-hot encoded levels were aggregated back to their original attributes to avoid reporting a separate importance value for every category. The resulting grouped feature importance is presented in Figure 8.



(a) Fuel consumption

(b) CO_2 emissions

Figure 8. Grouped SHAP feature importance obtained from the strongest standalone XGBoost models.

The most influential attributes for fuel-consumption prediction were engine size, vehicle class, and fuel type, whereas CO_2 predictions were primarily influenced by engine size, vehicle class, and transmission. These rankings are consistent with the engineering expectation that powertrain and structural vehicle characteristics influence energy demand and emissions. However, SHAP values represent associations learned by the models and should not be interpreted as causal effects.

Agreement between the measured and predicted values was examined using the scatter plots presented in Figure 9. The identity line represents perfect agreement and provides a visual reference for identifying systematic underprediction or overprediction.

The predictions were closely aligned with the identity line over the central portions of both target ranges. The largest deviations occurred at the upper end, where several high-consumption and high-emission vehicles were underpredicted. This indicates strong overall agreement but reduced accuracy for extreme vehicle configurations.

Residual diagnostics were subsequently used to examine systematic bias, error dispersion, and temporal consistency beyond the aggregate evaluation metrics. Figure 10 presents the residuals of the best valid ensemble, Ridge OOF stacking, as functions of the predicted values, as empirical distributions, and separately for each temporal test year.

The mean residuals were close to zero for both targets (0.034 L/100 km for fuel consumption and -0.071 g/km for CO_2), indicating little overall systematic bias. Nevertheless, residual dispersion increased for some high-consumption and high-emission vehicles, suggesting that the errors were not perfectly homoscedastic. The year-specific panels show only small changes in the mean residual from 2021 to 2024, although the uncertainty intervals widen in the later test years.

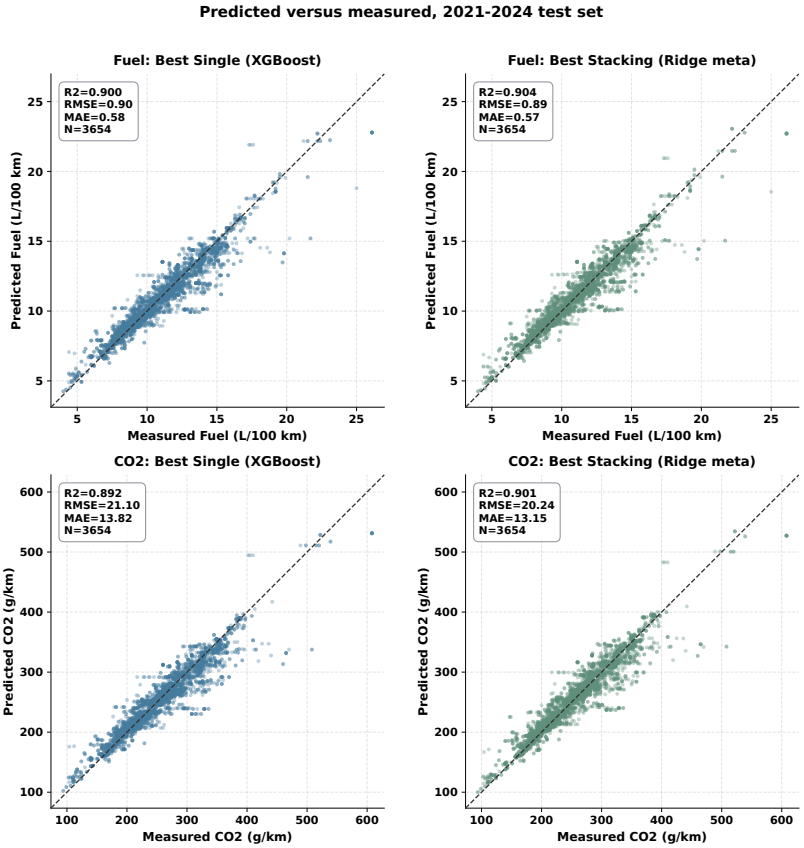


Figure 9. Measured versus predicted fuel-consumption and CO_2 values for the best individual model and best valid ensemble on the 2021–2024 temporal test set.

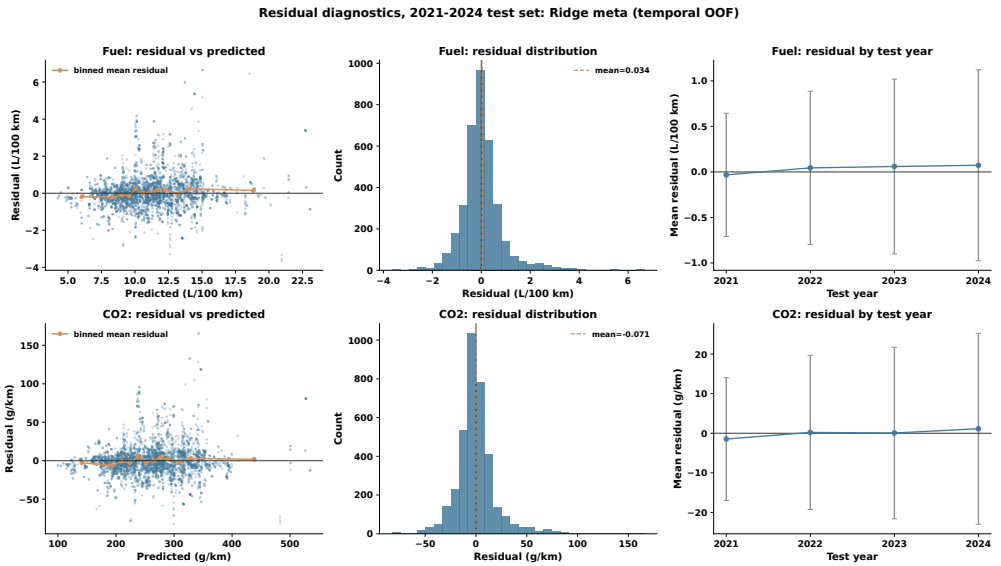


Figure 10. Residual diagnostics for Ridge temporal OOF stacking on the 2021–2024 test set. The panels show residuals versus predicted values, residual distributions, and mean residuals by test year for fuel consumption and CO_2 emissions.

To identify the vehicle configurations associated with the largest prediction errors, the results of the strongest standalone XGBoost models were further examined by vehicle class and fuel type. For fuel consumption, standard pickup trucks produced the largest class-specific error, with an RMSE of 1.2168 L/100 km ($n = 408$), followed by two-seaters with an RMSE of 1.1639 L/100 km ($n = 227$). Among the well-represented classes, mid-size vehicles obtained the lowest RMSE, at 0.6516 L/100 km ($n = 432$). Standard pickup trucks also showed a positive mean residual of 0.4409 L/100 km, indicating a tendency toward underprediction.

For CO_2 emissions, standard pickup trucks were again the most difficult class, with an RMSE of 28.8544 g/km and a mean residual of 10.7646 g/km. Two-seaters followed with an RMSE of 26.7047 g/km, whereas mid-size vehicles obtained a lower RMSE of 14.5289 g/km. The apparently low errors obtained for passenger vans should be interpreted cautiously because this category contained only four test records.

The fuel-type analysis also revealed heterogeneous error patterns. Ethanol E85 vehicles produced the largest fuel-consumption RMSE (1.0857 L/100 km, $n = 50$), whereas diesel vehicles obtained the lowest value (0.5336 L/100 km, $n = 95$). For CO_2 emissions, premium-gasoline vehicles produced the largest RMSE (21.7878 g/km, $n = 1874$), followed by regular-gasoline vehicles (20.6357 g/km, $n = 1635$). Because the category sizes differed, these group-level results should be interpreted together with their corresponding sample sizes.

The practical suitability of the evaluated models was examined by considering their computational requirements in addition to predictive accuracy. Table 9 reports final training time, inference latency per 1,000 rows, peak process memory, serialized model size, and execution device. All measurements were obtained on the same declared hardware, although XGBoost used the GPU and the remaining standalone models used the CPU. The reported training time represents final model fitting and does not include hyperparameter-search time.

Table 9. Computational requirements of the evaluated models, reported separately for each prediction target.

Target	Model	Device	Training (s)	Inference (ms/1,000 rows)	Peak memory (MB)	Size (MB)
Fuel	Linear Regression	CPU	2.578	36.74	775.84	13.81
Fuel	Random Forest	CPU	5.427	26.54	759.98	80.34
Fuel	Extra Trees	CPU	20.450	26.73	774.98	122.09
Fuel	SVR	CPU	19.156	2888.39	828.69	57.58
Fuel	LightGBM	CPU	10.336	33.91	796.31	7.89
Fuel	XGBoost	GPU	4.387	35.31	713.25	1.49
Fuel	CatBoost	CPU	15.232	2.56	568.79	3.80
Fuel	FT-Transformer	CPU	47.992	22.38	654.34	0.70
Fuel	Ridge OOF stack	GPU/CPU	177.426	63.19	848.79	2.32
CO_2	Linear Regression	CPU	2.406	36.88	792.30	13.81
CO_2	Random Forest	CPU	5.118	26.48	745.61	84.50
CO_2	Extra Trees	CPU	27.437	50.51	808.55	130.23
CO_2	SVR	CPU	7.688	1504.55	813.82	30.06
CO_2	LightGBM	CPU	10.822	33.73	843.05	7.84
CO_2	XGBoost	GPU	4.119	38.92	715.48	1.38
CO_2	CatBoost	CPU	17.831	2.57	560.21	3.92
CO_2	FT-Transformer	CPU	68.245	39.91	652.28	1.21
CO_2	Ridge OOF stack	GPU/CPU	246.081	84.38	856.77	2.72

Linear Regression required the shortest training time, whereas CatBoost provided the lowest inference latency. SVR was by far the slowest model during inference, while Ridge OOF stacking required the largest overall training time and peak memory. XGBoost provided the most favorable standalone accuracy–cost compromise: it achieved the best individual temporal accuracy, required approximately four seconds of GPU training, and produced models smaller than 1.5 MB. Ridge stacking improved predictive accuracy but incurred substantially higher computational cost.

Previous studies derived from the NRCan Fuel Consumption Ratings database have reported high predictive performance. However, their results were obtained using different data periods, predictors, target formulations, and validation protocols. The methodological comparison with the most closely related NRCan-based studies is presented in Table 10.

Table 10. Comparison with related studies derived from the Natural Resources Canada Fuel Consumption Ratings data.

Study	Dataset version	Target and best model	Input setting	Evaluation protocol	Reported R^2
Hien and Kor [31]	4,973 records, 2017–2021	Fuel: multiple linear regression; CO_2: degree-5 polynomial regression	City and highway consumption for fuel prediction; total fuel consumption for CO_2 prediction	Random 80/20 hold-out; no temporal test	Fuel: 0.9997; CO_2: 0.9122
Natarajan *et al.* [35]	7,384 records, 2017–2021	CO_2: CatBoost	City, highway, and combined fuel-consumption variables included	Random 80/20 hold-out; no temporal test	0.9960
Nurdin *et al.* [36]	9,185 records, 2015–2023; 6,951 retained after preprocessing; 764 records for 2024	Combined fuel consumption: Random Forest	City, highway, combined mpg, and CO_2-related variables included	Random 80/20 split with GridSearchCV; separate 2024 test	0.8845 (2024 test)
Guo *et al.* [37]	26,075 records, 1995–2022	CO_2: Gradient Boosting Machine	City, highway, and combined fuel-consumption variables included	Random 80/20 hold-out with five-fold cross-validation	0.9973
Satya *et al.* [38]	7,385 records; model years not reported	CO_2: tuned Random Forest	Combined fuel consumption included as the dominant predictor	Random train/test split; five-fold GridSearchCV	0.9982
Our Approach	**10,060 records, 2015–2024**	**Fuel: Ridge OOF stacking; CO_2: Ridge OOF stacking**	**Vehicle specifications only; no target-derived or cross-target predictors**	**Validation within 2015–2020; untouched temporal test on 2021–2024**	**Fuel: 0.9042; CO_2: 0.9007**

Note: All studies use data derived from the NRCan Fuel Consumption Ratings database, but they differ in data release, sample period, predictor set, target formulation, and validation protocol. Several previous studies used measured fuel-consumption variables to predict CO_2 emissions. Some fuel-consumption models also used the city and highway components from which combined consumption is calculated. In contrast, the present study excludes target-derived and cross-target predictors. The reported R^2 values therefore provide methodological context rather than a direct ranking of predictive performance.

Among the related studies, Nurdin et al. [36] provide the closest temporal comparison because their model was additionally evaluated using future-year data. Nevertheless, their training period, temporal boundary, and predictor configuration differ from those used in the present study. Other studies included city, highway, or combined fuel-consumption measurements as predictors of CO_2 emissions, resulting in a less restrictive prediction task.

In contrast, the present study excludes target-derived and cross-target predictors, performs model selection exclusively within the 2015–2020 training period, and reserves the complete 2021–2024 period for final evaluation. Under this protocol, Ridge OOF stacking achieved the strongest overall fuel-consumption performance, with $R^2 = 0.9042$, and the strongest CO_2 performance, with $R^2 = 0.9007$. It improved upon the strongest standalone XGBoost model for both targets. For CO_2, however, its RMSE was practically identical to that obtained by simple prediction averaging.

Taken together, the temporal, statistical, external, interpretability, and computational analyses demonstrate that model quality cannot be assessed using a single accuracy value or a random train–test split. Reliable vehicle-level prediction requires leakage control, temporally ordered validation, explicit uncertainty assessment, cross-dataset evaluation, and consideration of computational cost. These elements constitute the principal methodological contribution of the present study.

5. Conclusion And Future Work

This study presented a leakage-controlled comparative framework for predicting vehicle fuel consumption and CO_2 emissions from technical vehicle specifications. Eight individual regression models and several stacking configurations were evaluated using Natural Resources Canada records covering 2015–2024. Model development and selection were restricted to the 2015–2020 period, while records from 2021–2024 were retained as an

untouched temporal test set. Among the individual models, XGBoost achieved the strongest temporal-test performance, with an RMSE of 0.9032 L/100 km and $R^2 = 0.9003$ for fuel consumption, and an RMSE of 21.0965 g/km and $R^2 = 0.8921$ for CO_2 emissions. Among all valid configurations, Ridge OOF stacking achieved the lowest RMSE for both targets: 0.8850 L/100 km for fuel consumption and 20.2415 g/km for CO_2, corresponding to R^2 values of 0.9042 and 0.9007, respectively. These results were obtained without using target-derived or cross-target predictors.

The experiments further indicate that PCA should be selected separately for each algorithm and target rather than imposed uniformly. Valid temporal stacking improved upon the strongest individual model for both tasks, although its advantage over simple averaging was negligible for CO_2 emissions. The year-by-year results revealed progressive degradation from 2021 to 2024, while the EPA evaluation showed that the combined geographic and future-year shift was more demanding than matched-year transfer. These findings emphasize the importance of leakage control, temporally ordered validation, statistical uncertainty assessment, and cross-dataset testing rather than model selection based solely on architectural complexity.

The findings remain subject to several limitations. The NRCan dataset contains standardized certification ratings rather than unrestricted on-road measurements and therefore does not represent driving behavior, traffic conditions, road gradient, weather, vehicle load, or maintenance status. External evaluation using U.S. data may also be affected by differences in vehicle specifications and regulatory testing procedures. Moreover, fuel consumption and CO_2 emissions were modeled as separate regression tasks, and predictive uncertainty was not explicitly quantified. Future research should consequently examine on-road measurements, additional countries, heavy-duty vehicles, and electrified powertrains. Further extensions may include uncertainty-aware prediction, multi-task learning, domain adaptation across regulatory environments, and model calibration under changing vehicle technologies and operating conditions.

Data and Code Availability

The primary NRCan dataset is publicly available through the Government of Canada Open Government Portal, while the external vehicle records are available from the U.S. EPA Fuel Economy database. The complete source code, final executable notebook, pinned dependency file, execution instructions, and scripts used to generate the reported tables and figures are publicly available at:https://github.com/dohaaa/SOIC-Elkalef.git.

REFERENCES

1. E. S. Soegoto, R. D. Utami, and Y. A. Hermawan, "Influence of artificial intelligence in automotive industry," *J. Phys.: Conf. Ser.*, vol. 1402, no. 6, Dec. 2019, doi: 10.1088/1742-6596/1402/6/066081.
2. Ş. Yıldırım, A. D. Yucekaya, M. Hekimoğlu, M. Ucal, M. N. Aydin, and İ. Kalafat, "AI-driven predictive maintenance for workforce and service optimization in the automotive sector," *Applied Sciences*, vol. 15, no. 11, Art. no. 6282, 2025, doi: 10.3390/app15116282.
3. J. M. Shah, N. A. Natraj, G. G. Hallur, and A. Aslekar, "Artificial Intelligence (AI) in the Automotive Industry and the use of Exoskeletons in the Manufacturing Sector of the Automotive Industry," in *Proc. 2023 Int. Conf. Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE, Mar. 2023, pp. 428–432, doi: 10.1109/ICSCDS56580.2023.10105009.
4. M. Janjic, *The Role of the International Energy Agency (IEA) in Global Energy Governance*, Master's thesis, University of Piraeus, 2025.
5. D. H. A. Wardhana and M. R. Prawira, "The Analysis of Indonesia's Climate Change Policies in Response to the 2021 Intergovernmental Panel on Climate Change (IPCC) Assessment Report/AR6 Group 1 (2021–2023)," in *Proc. Int. Relations Indones. Foreign Policy Conf. (IROFONIC)*, vol. 1, no. 1, pp. 42–53, 2024.
6. P. Jaramillo et al., "Transport (Chapter 10)," in *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2023, pp. 1049–1160.
7. S. Solaymani and J. Botero, "Reducing carbon emissions from transport sector: Experience and policy design considerations," *Sustainability*, vol. 17, no. 9, Art. no. 3762, 2025.
8. International Energy Agency (IEA), *Greenhouse Gas Emissions from Energy: Highlights*, Paris, France: IEA, 2024.
9. M. A. Hamed, M. H. Khafagy, and R. M. Badry, "Fuel Consumption Prediction Model using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, 2021, doi: 10.14569/IJACSA.2021.0121146.
10. R. B. Kale and N. F. Shaikh, "Analysis of Learning Algorithms for Predicting Carbon Emissions of Light-Duty Vehicles," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 7, 2024, doi: 10.14569/IJACSA.2024.0150757.

11. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, ACM, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
12. A. Mackiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.*, vol. 19, no. 3, pp. 303–342, Mar. 1993, doi: 10.1016/0098-3004(93)90090-R.
13. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 18932–18943, 2021.
14. S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
15. S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," *arXiv preprint arXiv:1909.06312*, 2019.
16. X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.
17. F. D. B. Albuquerque et al., "Greenhouse gas emissions associated with road transport projects: Current status, benchmarking, and assessment tools," *Transp. Res. Procedia*, vol. 48, pp. 2018–2030, 2020, doi: 10.1016/j.trpro.2020.08.261.
18. A. A. Alhussan, M. Metwally, and S. K. Towfek, "Predicting CO_2 Emissions with Advanced Deep Learning Models and a Hybrid Greylag Goose Optimization Algorithm," *Mathematics*, vol. 13, no. 9, Art. no. 1481, Apr. 2025, doi: 10.3390/math13091481.
19. S. Zahid and U. Jamil, "Data-driven machine learning techniques for fuel economy prediction in sustainable transportation systems," *Green Energy and Intelligent Transportation*, vol. 4, no. 2, Art. no. 100303, 2025, doi: 10.1016/j.geits.2025.100303.
20. J. Udoh, J. Lu, and Q. Xu, "Application of Machine Learning to Predict CO_2 Emissions in Light-Duty Vehicles," *Sensors*, vol. 24, no. 24, Art. no. 8219, Dec. 2024, doi: 10.3390/s24248219.
21. D. R., H. K. Joy, S. R., E. A. J. A., and V. D., "Machine Learning and Deep Learning Analysis of Vehicle Carbon Footprint," *Int. J. Environ. Impacts*, vol. 7, no. 2, pp. 287–292, Jun. 2024, doi: 10.18280/ijei.070213.
22. M. Ehsani, A. Ahmadi, and D. Fadai, "Modeling of vehicle fuel consumption and carbon dioxide emission in road transport," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 1638–1648, Jan. 2016, doi: 10.1016/j.rser.2015.08.062.
23. E. Vyshka, S. Sefa, F. Basholli, and D. Lumi, "Analysis of the pollutant emissions from vehicles and their impact on the environment and public health," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27660–27671, Oct. 2025, doi: 10.48084/etasr.12747.
24. M. Hamoumi, M. Benhadou, and A. Haddout, "A holistic modeling framework for operational excellence in the automotive industry: An integrated perspective on technical systems and human dynamics," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27570–27580, Oct. 2025, doi: 10.48084/etasr.13521.
25. M. Chaman, A. El Maliki, N. Jariri, A. El Mrabet, H. Dahou, H. Laamari, and A. Hadjoudja, "A real-time vehicle detection system for ADAS in autonomous vehicles using YOLOv11 deep neural network on embedded edge platforms," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 28077–28082, 2025, doi: 10.48084/etasr.12138.
26. Natural Resources Canada, "2015–2024 Fuel Consumption Ratings," Government of Canada Open Government Portal, Mar. 2026. [Online]. Available: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64/resource/c98b9dc8-b23f-4cd8-8b19-e892da1e4688.
27. T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.
28. D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/peerj-cs.623.
29. A. H. Al-Nefaie and T. H. H. Aldhyani, "Predicting CO_2 emissions from traffic vehicles for a sustainable and smart environment using deep learning," *Sustainability*, vol. 15, no. 9, Art. no. 7615, 2023, doi: 10.3390/su15097615.
30. G. M. I. Alam, S. A. Tanim, S. K. Sarker, Y. Watanobe, R. Islam, M. F. Mridha, and K. Nur, "Deep learning model based prediction of vehicle CO_2 emissions with eXplainable AI integration for sustainable environment," *Scientific Reports*, vol. 15, Art. no. 3655, Jan. 2025, doi: 10.1038/s41598-025-87233-y.
31. N. L. H. Hien and A.-L. Kor, "Analysis and prediction model of fuel consumption and carbon dioxide emissions of light-duty vehicles," *Applied Sciences*, vol. 12, no. 2, Art. no. 803, 2022, doi: 10.3390/app12020803.
32. H. Pinto Vega, L. F. Grisales-Noreña, and V. Botero-Gómez, "Optimizing energy management in AC microgrids: A comparative study of metaheuristic algorithms for minimizing energy losses and CO_2 emissions," *Statistics, Optimization & Information Computing*, vol. 14, no. 2, pp. 636–662, 2025, doi: 10.19139/soic-2310-5070-2455.
33. A. Abdelsamiea, E. Elashkar, M. A. Zayed, and M. F. Abd El-Aal, "Assessing the relationship between global health spending and carbon emissions using a gradient-boosting algorithm," *Statistics, Optimization & Information Computing*, vol. 15, no. 1, pp. 128–137, 2025, doi: 10.19139/soic-2310-5070-2858.
34. K. Boumais and F. Messaoudi, "Short-term load forecasting method for renewable energy integration and grid stability using CNN, LSTM, and Transformer models," *Statistics, Optimization & Information Computing*, vol. 13, no. 4, pp. 1578–1594, 2024, doi: 10.19139/soic-2310-5070-2256.
35. Y. Natarajan, G. Wadhwa, K. R. Sri Preethaa, and A. Paul, "Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms," *Electronics*, vol. 12, no. 10, Art. no. 2288, 2023, doi: 10.3390/electronics12102288.
36. M. Nurdin, Wamiliana, A. Junaidi, and F. R. Lumbanraja, "Comparison of support vector regression and random forest regression performance in vehicle fuel consumption prediction," *Integra: Journal of Integrated Mathematics and Computer Science*, vol. 1, no. 2, pp. 60–67, 2024, doi: 10.26554/integrajimcs.20241221.
37. X. Guo, R. Kou, and X. He, "Towards carbon neutrality: Machine learning analysis of vehicle emissions in Canada," *Sustainability*, vol. 16, no. 23, Art. no. 10526, 2024, doi: 10.3390/su162310526.
38. B. Satya, M. D. Miqoilla, A. Aminuddin, A. N. L. Nabhaan, M. B. A. Miah, and H. R. Enriquez, "Machine learning prediction of CO_2 emissions from light-duty vehicles in Canada," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, pp. 32260–32266, Feb. 2026, doi: 10.48084/etasr.15599.