International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 2, June 2026, pp. 501~517
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i2.pp501-517

Estimating road transport emissions in Thailand using data-driven modeling approaches

Phachaya Chaiwchan[1], Kunyanuth Kularbphettong[2]
[1]Digital Innovation Management Program, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand
[2]Computer Science Program, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand

Article Info

Article history:

Received Apr 10, 2025
Revised Mar 13, 2026
Accepted Apr 22, 2026

Keywords:

Data-driven modeling
Emission estimation
Machine learning
Urban pollution
Vehicle CO_2 emissions

ABSTRACT

Traffic emissions remain a pivotal factor in Bangkok's air quality issues, particularly as the city expands and vehicle types become more varied and traditional emission-inventory tools, such as Computer Programme to Calculate Emissions from Road Transport (COPERT) and laboratory driving cycles, are useful but often fall short when applied to congested real-world conditions. This work responds to that gap by combining laboratory measurements, regulatory emission factors, and vehicle-registration information to estimate CO_2 emissions at the vehicle level. Data from 1,846,111 registered vehicles (January 2023–January 2025) were consolidated from Environmental Protection Agency (EPA) fuel-economy tests, COPERT/Euro standards, and the Thai DLT registry. After harmonizing attributes across all sources, additional features were created and screened using recursive feature elimination (RFE). Three models were trained and evaluated using coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). CatBoost provided the most reliable predictions (RMSE =23.4 g/km; MAE =14.6 g/km; R^2 =0.921), followed by deep neural network (DNN) (R^2 =0.887) and support vector regression (SVR) (R^2 =0.846). RFE consistently identified fuel-consumption measures, emission factors, and tailpipe CO_2 as the most influential variables and the results provide a clearer picture of how vehicle technologies contribute to Bangkok's CO_2 burden and support policy measures such as emission taxes, incentives for low-emission vehicles, and monitoring of high-emission groups.

Corresponding Author:

Kunyanuth Kularbphettong
Computer Science Program, Faculty of Science and Technology, Suan Sunandha Rajabhat University
1 U-Thong Nok Road, Bangkok 10300, Thailand
Email: kunyanuth.ku@ssru.ac.th

1. INTRODUCTION

The transportation sector remains one of the main sources of greenhouse gas emissions in Thailand, and CO^2 from vehicle exhaust continues to drive national warming trends at a noticeable scale [1]–[4]. Air pollution linked to traffic also affects public health directly. The impacts are particularly serious for groups already at risk, including children, older adults, and people with chronic respiratory or cardiovascular illnesses.

World Health Organization (WHO) reports that outdoor air pollution accounted for roughly 68% of premature deaths in 2019, mainly from ischemic heart disease, stroke, chronic obstructive pulmonary disease (COPD), lower respiratory infections, and lung cancer [5]. As of September 2023, Thailand had more than 44 million registered vehicles. When combined with persistent congestion, stop-and-go driving, and mixed vehicle technologies, these conditions make CO_2 emissions in Bangkok far more variable than national averages would suggest [6].

Such circumstances point to the need for more detailed and data-driven emission-estimation approaches to support urban planning and evidence-based policy. The recent adoption of Euro 5 and the transition toward Euro 6 standards reflect Thailand's efforts to reduce pollutants such as NOx, $PM_{2.5}$, hydrocarbons, and CO. However, these regulations provide relatively limited guidance on real-world CO_2 emissions [7].

Previous research in Thailand has mainly relied on deterministic tools, particularly Computer Programme to Calculate Emissions from Road Transport (COPERT) and Euro 4–6 emission factors, which estimate emissions based on standardized driving cycles and typical fuel characteristics [8], [9]. While these methods are widely used, they struggle to capture the complexity of real driving in metropolitan areas. The research in [8]–[10] shows that COPERT outputs can vary considerably and sometimes overpredict CO_2 emissions at low urban speeds, indicating the need for adjustment using local data and, in some cases, the value of integrating machine learning techniques. Emission factors unique to Bangkok that are generated from chassis dynamometer testing have also been established to improve accuracy under local driving conditions [11], even though their practical application is still limited because of laboratory constraints and small sample quantities. Further analysis with COPERT for the Bangkok Metropolitan Region provides context for urban emissions [12]; nevertheless, these methodologies still lack resolution at the vehicle level.

Internationally, machine learning models have demonstrated the ability to outperform traditional regression for CO_2 estimation, including explainable machine learning frameworks [13], multi-model ensembles comparing up to 18 algorithms [14], and applications highlighting the superiority of machine learning over classical statistical methods in transport-emission analysis [15]. Moreover, CatBoost-based forecasting has shown strong accuracy for light-duty vehicle CO_2 prediction and robustness in heterogeneous fleets similar to Bangkok's [16]. Broader reviews of machine learning, artificial neural networks, and deep learning methods in transport-emissions modeling further support the growing shift toward data-driven approaches [17]. However, most existing machine learning studies rely on certification datasets, small experimental samples, or vehicle fleets from non-Asian regions, limiting their applicability to complex, mixed-technology megacities such as Bangkok.

Table 1 presents a comparative overview of conventional emission estimation frameworks and machine learning approaches frequently applied in transportation-related CO_2 emission studies. The comparison highlights their methodological strengths and practical limitations. It also addresses their relevance to Bangkok's urban traffic conditions and heterogeneous vehicle registration characteristics.

These collective findings reveal two essential gaps. First, few studies integrate detailed vehicle attributes, such as model year, engine displacement, fuel type, and Euro certification, with standardized emission factors to enable vehicle-level CO_2 estimation. Second, to the best of current knowledge, no prior research has combined EPA laboratory-tested CO_2 values, COPERT/Euro regulatory factors, and a large-scale real-world registration dataset within a unified analytical and machine learning framework for Bangkok.

To address these gaps, this study integrates three complementary datasets:

i) Laboratory-based fuel economy and CO_2 measurements from the U.S. Environmental Protection Agency (EPA).
ii) A comprehensive registry of 1.846 million vehicles recorded by the Department of Land Transport (DLT) in Bangkok between 2023 and 2025.
iii) COPERT/Euro 4–6 emission factors commonly used in international emission-inventory systems.

These sources are harmonized into a unified vehicle-level analytical dataset to develop and evaluate three machine learning models: CatBoost, support vector regression (SVR), and deep neural networks (DNN). The models are used to estimate real-world CO_2 emissions at the individual-vehicle level and to compare predictive performance across ensemble-based, kernel-driven, and deep learning approaches. In addition, the analysis examines how key vehicle characteristics, such as engine displacement, fuel type, and Euro certification, shape CO_2-emission behavior, thereby offering empirical insights that may inform targeted regulatory measures and future transport-emission policies.

The remainder of the article is organized as follows. Section 2 provides a detailed description of the study area, the data sources used, and the preprocessing steps applied to prepare the data for analysis. Section 3 describes the experimental workflow and the procedures followed during model training. Section 4 presents the results, including assessments of variable importance, measures of prediction accuracy, and an analysis of emission characteristics. Section 5 concludes the paper with policy considerations and suggestions for directions for future work.

Table 1. Comparative overview of CO_2 emission estimation methods

Approach/model	Strengths	Limitations	Relevance to Bangkok context	Key references
Thai local emission studies	Provide localized emission factors and real-world contextual insights	Limited scalability for large fleet-level prediction and dynamic forecasting	Useful for regional calibration and validation of Bangkok-specific emission characteristics	Nilrit *et al*. [2], [3], Sirithian *et al.* [11], Kuson *et al*. [12]
COPERT	Consistent, internationally recognized, and suitable for national inventories	May overestimate emissions under low-speed congestion and stop-and-go traffic conditions	Bangkok's highly congested urban traffic requires empirical adjustment and localized calibration	Lejri *et al*. [8], Dey *et al*. [9], COPERT/EEA [18], Kholod *et al*. [19], Ntziachristos *et al*. [20]
Ensemble machine learning (Random forest/XGBoost)	Robust to outliers; effective for nonlinear tabular prediction and feature interaction modeling	Requires careful encoding and may suffer from leakage in high-cardinality categorical variables	Performs reasonably well but may struggle with Bangkok's extremely heterogeneous DLT categorical structure	Gurcan [14], Çınarer *et al*. [15], Zhou *et al*. [16]
SVR	Strong nonlinear approximation capability and effective generalization performance	Sensitive to kernel selection and hyperparameter tuning; computationally expensive for very large datasets	Bangkok's dataset (~1.8M) poses computational challenges	Çınarer *et al.* [15], Natarajan *et al.* [21]
Open vehicle specification and fuel economy datasets	Large-scale standardized vehicle information; useful for feature engineering and fleet matching	May lack localized driving behavior and operational variability	Supports integration with Bangkok DLT registration attributes and fleet heterogeneity	Opendatasoft dataset [22], DLT dataset [23]
Euro 4–6 certification standards	Enables standardized cross-manufacturer comparison and regulatory benchmarking	Laboratory-based values may differ substantially from real-world urban emissions	Bangkok fleet includes many pre-Euro vehicles and mixed technologies	European Commission [24], Fontaras *et al*. [25]
Simulation models (PHEM, MOVES)	Captures dynamic behavior; high realism for micro-scale analysis	Requires high-resolution speed, time, and trajectory data	Bangkok lacks high-resolution speed profile data	Lejri *et al*. [8], Fontaras *et al*. [25]
Linear/traditional regression	Simple; interpretable; reproducible	Limited ability to model nonlinear interactions and heterogeneous fleet behavior	Bangkok's mixed vehicle technologies and nonlinear traffic conditions reduce predictive capability	Natarajan *et al.* [21]
Gradient boosting for categorical data (CatBoost)	Efficiently handles high-cardinality categorical variables; reduces target leakage; strong performance on heterogeneous tabular datasets	Requires hyperparameter tuning and computational optimization for large-scale training	Highly suitable for Bangkok's diverse fleet composition and large categorical feature space	Sun *et al.* [26], Feng *et al.* [27]
DNN	Captures hierarchical nonlinear relationships and complex feature interactions	Sensitive to overfitting; requires normalization; high computational cost	Effective for large-scale prediction but potentially unstable under mixed ICE/EV/hybrid fleet distributions	Ghorbal *et al.* [28], Komara *et al*. [29], Gago *et al.* [30], Al-Nefaie and Aldhyani [31]

2. MATERIALS AND METHOD

2.1. Study area and data sources

This study focuses on Bangkok, Thailand's capital and largest city, with over 10 million population and many urban subdistricts. Figure 1 shows significant business zones, high-traffic routes, and mixed residential and commercial districts in the specified area. The diversified vehicle fleet in Bangkok, including passenger vehicles, buses, heavy-duty trucks, motorcycles, and others, provides a realistic environment for measuring CO_2 emissions. The incorporation of numerous subdistricts facilitates variability in road infrastructure, traffic density, and vehicle usage patterns, all of which are critical determinants in emission modeling. To estimate CO_2 emissions from vehicles registered in Bangkok, Thailand, this study employs three primary data sources to construct machine learning models.

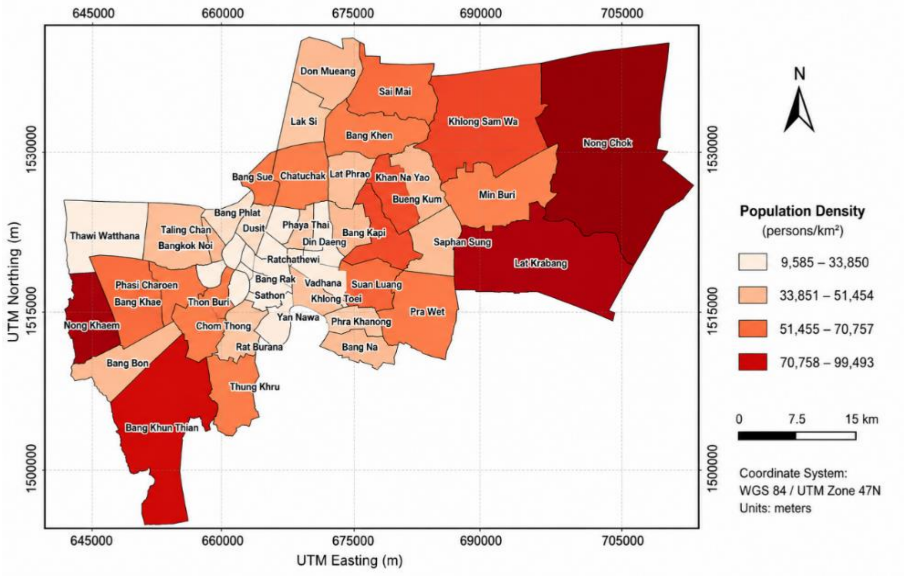

Figure 1. Study area map

2.1.1. Environmental Protection Agency vehicle emissions and fuel economy dataset

The initial dataset was sourced from the U.S. EPA [22], [25], comprising standardized laboratory-tested data on vehicle fuel economy, engine displacement, manufacturer specifications, and CO_2 emissions measured in grams per kilometer (g/km). The measurements produced at the EPA National Vehicle and Fuel Emissions Laboratory function as baseline reference values. They also offer dependable estimates of vehicle-level environmental performance.

2.1.2. Bangkok vehicle registration dataset

The second dataset comprises 1,846,111 vehicle registration records gathered by Thailand's DLT from January 2023 to January 2025 [23]. The dataset encompasses vehicle size class, brand, model, fuel type, engine type, and emission-control technologies. Thereby offering a precise representation of Bangkok's on-road vehicle fleet for localized emission modeling and trend analysis.

2.1.3. Emission standards and mapping variables

The third dataset includes CO_2 emission factors obtained from COPERT 5 and Euro 4–6 emission certification standards, as reported by the European Environment Agency [18]–[20], [24], [32]. Emission factors were allocated to each registered vehicle based on criteria including certification level, fuel type, and engine technology. In the absence of direct emission numbers from EPA records, COPERT-based estimates were utilized, and this correlation facilitates a standardized comparison of measured and regulatory emission values and supports the prediction of CO_2 emissions across various vehicle classes.

2.2. Data preprocessing

Real-world vehicle emissions differ by 30%–40% from laboratory test values due to driving behavior, traffic conditions, and environmental factors not represented in controlled testing [25]. The implementation of a structured data pretreatment and integration procedure was necessary to resolve these disparities. It also enhances the realism of emission estimation for Bangkok.

2.2.1. Data cleaning and harmonization

All datasets were standardized to ensure consistency in variable names, units, and categorical formats. Manufacturer and model fields were normalized across EPA and DLT records. Invalid, duplicated, or incomplete entries were removed. Vehicles that could not be associated with emission factors, such as non-motorized or special-purpose vehicles were excluded from the analysis.

2.2.2. Integration of EPA, DLT, and COPERT/Euro data

A multi-stage integration procedure was applied to combine the three primary data sources: i) EPA laboratory-tested CO_2 emissions and vehicle specifications; ii) DLT vehicle registration records for

Bangkok from January 2023 to January 2025; and iii) COPERT and Euro 4–6 emission factors. Every registered vehicle was correlated with EPA information according to brand, model, engine displacement, and fuel type. In the absence of laboratory-measured CO_2 data, coefficients derived from the COPERT methodology were allocated based on vehicle classification, fuel type, and Euro certification. This methodology guaranteed that every 1,846,111 vehicles were allocated a designated CO_2 reference value for the construction of the machine learning model.

From the full 1.846 million vehicles, 717,089 vehicles (≈38.8%) could be matched with valid emission information (EPA or COPERT). Table 2 summarizes the characteristics of the resulting integrated dataset. This empirical summary corresponds with documented vehicle trends in Bangkok, illustrating a fleet predominantly composed of gasoline passenger vehicles, alongside an increasing prevalence of battery electric vehicles (BEV) and late-Euro-standard automobiles.

Table 2. Summary statistics of the integrated Bangkok fleet dataset

Variable	Value/distribution
Total matched vehicles	717,089
Vehicle class (%)	Passenger cars 94.5%, motorcycles 2.8%, pickups/SUVs 1.8%, and others 0.9%
Fuel type (%)	Gasoline 93.7%, BEV 3.7%, diesel 1.6%, and hybrid/PHEV 1.0%
Engine displacement (cc)	Mean±SD =2,057±5,714
Euro standard distribution	Euro ≤3 (37.7%), Euro 4 (38.3%), Euro 5 (12.3%), and Euro 6 (11.4%)
EPA CO_2 (g/km)	Mean±SD =322.2±55.3
Registration year range	1990–2025

2.2.3. Feature engineering

Several derived variables were created to improve predictive capacity of machine learning models. Engine displacement was grouped into practical categories (e.g., <1.2 L, 1.2–1.8 L, 1.8–2.5 L, and >2.5 L) to capture nonlinear relationships between engine size and CO_2 output. Euro emission standards were assigned based on regulatory implementation timelines and vehicle manufacturing years, while vehicle classes, like passenger cars, pickup trucks, motorcycles, electric vehicles, and sport-utility vehicles, were categorized to reflect functional differences in real-world operation. Fuel system configurations were further classified into internal combustion engine (ICE), hybrid, plug-in hybrid (PHEV), and BEV types. A standardized estimate of annual mileage was applied across vehicles to support subsequent aggregation and comparative analysis. Categorical features were processed according to the requirements of each modeling technique: CatBoost was able to ingest raw categorical variables directly using its native encoding mechanisms, whereas SVR and DNN required numerical transformation via one-hot encoding or embedding representations.

2.2.4. Cross-validation strategy

To enhance robustness and reduce sample bias, a 5-fold cross-validation method was employed instead of a static train–validation–test division and the dataset was randomly partitioned into five equal subsets, with each fold utilized repeatedly as the testing set while the remaining four functioned as the training set. Performance metrics, including coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), were calculated for each repeat, with the mean and standard deviation subsequently reported. Hence, this cross-validation strategy ensures that model evaluation remains stable and representative across diverse vehicle types, fuel categories, and emission levels.

2.2.5. Model input preparation

The machine learning models were trained utilizing vehicle-level predictors, including engine displacement, fuel consumption, fuel type, vehicle class, and Euro emission standard. Before final model building, recursive feature elimination (RFE) was utilized to ascertain the most significant predictors and minimize redundancy in the feature set. This comprehensive preprocessing pipeline, incorporating regulatory parameters, empirically validated emission values, and actual fleet characteristics, yielded a high-resolution dataset appropriate for supervised learning. The curated dataset offers a solid basis for precisely modeling vehicle-level CO_2 emissions under various traffic and vehicle conditions in Bangkok's urban setting.

2.3. Modeling techniques

Three supervised machine learning models were used to estimate vehicle-level CO_2 emissions using heterogeneous urban fleet data: DNN, SVR, and CatBoost. These models were chosen to overcome the shortcomings of deterministic emission models. They address their complementary strengths, which include handling categorical features, modeling nonlinear interactions, and capturing high-complexity patterns.

2.3.1. CatBoost

CatBoost is a gradient-boosting method specifically designed for tabular datasets that excels in handling categorical variables without necessitating considerable preparation. It builds an ensemble of decision trees in successive phases, with each tree rectifying the residual errors of its predecessors, so achieving elevated forecast accuracy. CatBoost is thus ideally suited for heterogeneous vehicle data, including brand, model, fuel type, Euro emission standard, and manufacture year, frequently seen in emission modeling research [25].

A major benefit of CatBoost its ordered target-based encoding, an approach that prevents target leakage by ensuring that categorical encodings are generated entirely from information available prior to each observation and this method substantially reduces overfitting, a prevalent problem in gradient boosting when applied to high-cardinality categorical features. CatBoost additionally incorporates ordered boosting, a variant of conventional boosting that significantly reduces prediction bias. The CatBoost prediction adheres to the standard gradient boosting framework, articulated as (1).

$$\hat{y}(x) = F_0(x) + \sum_{t=1}^{T} \eta_t h_t(x) \tag{1}$$

Where $\hat{y}_{CO_2}$ is the predicted CO_2 emissions, $F_0(x)$ is an initial model prediction (mean of CO_2), $f_t(x)$ is the prediction from the t^{th} decision tree, η_t is the learning rate, and T is the total number of trees.

This work involved the tuning of CatBoost by adjusting critical hyperparameters such as tree depth, iteration count, and learning rate. Regularization techniques, specifically ordered boosting, were employed to reduce overfitting. These techniques also improved generalizability across Bangkok's varied vehicle fleet.

2.3.2. Support vector regression

SVR, an extension of the support vector machine framework, was included as a baseline regression model to benchmark traditional machine learning performance against ensemble (CatBoost) and deep learning (DNN) methods. Using the radial basis function (RBF) kernel, SVR can effectively model intricate nonlinear associations across vehicle-level predictors, including engine size, fuel type, Euro standard, and fuel consumption. However, SVR is known to be sensitive to hyperparameter selection and may become computationally expensive when applied to large-scale datasets such as Bangkok's 1.84 million–vehicle registry.

The goal of SVR is to learn a function $f(x)$ that deviates from the true target values y by no more than a tolerance margin ε, while maintaining the function as flat as possible. This yields the ε–insensitive regression function, defined as (2).

$$\hat{y}(x) = w \cdot x + b \tag{2}$$

Where $\hat{y}(x)$ is the predicted CO_2 emission value, w is the weight vector learned by the model, x is the feature vector, which may include fuel type, Euro standard, miles per gallon (MPG), engine size, vehicle type, and other minus substances, and b is the bias term. SVR uses the ε–insensitive loss function, defined as (3).

$$L_\varepsilon(y, \hat{y}) = \begin{cases} 0, & if|y-\hat{y}| \le \varepsilon \\ |y - \hat{y}| - \varepsilon, & otherwise \end{cases} \tag{3}$$

By disregarding errors lower than ε, this formulation enables the model to concentrate on notable deviations instead of indiscriminately minimizing all training errors. Hyperparameters C, γ, and ε were optimized using k-fold cross-validation to mitigate overfitting and ensure robust performance across diverse vehicle categories. Although SVR provides a rigorous theoretical foundation, its scalability limitations and sensitivity to kernel parameters make it less effective for very large heterogeneous datasets, which is reflected in its comparatively lower predictive accuracy in this study.

2.3.3. Deep neural network

A feedforward DNN was implemented to model the complex nonlinear interactions between vehicle characteristics and CO_2 emission behavior. Although DNNs are traditionally applied to classification problems, they also demonstrate strong performance in regression tasks. In these tasks, the objective is to predict continuous numerical values such as vehicle-level CO_2 emissions derived from engine, fuel, and regulatory attributes.

The forward propagation process for each layer l in the network is defined as (4).

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \tag{4}$$

Where $z^{(l)}$ is the output of layer l before applying the activation function (pre-activation value), $W^{(l)}$ represents the weight vector of the connections in layer l, $a^{(l-1)}$ is the output (activation) from the previous layer (l−1), and $b^{(l)}$ denotes the bias term in layer l. Activation is then computed as (5).

$$a^{(l)} = \sigma\left(z^{(l)}\right) \tag{5}$$

The DNN employed the rectified linear unit (ReLU) as the activation function because of its stability and efficiency in deep architectures. Model optimization utilized the Adam optimizer, which adjusts learning rates throughout training to enhance convergence speed. To mitigate overfitting, particularly crucial for extensive and diverse datasets, regularization approaches were employed, including dropout layers and early stopping. The network's depth, neuron count per layer, and dropout rates were optimized by cross-validation on the training dataset, and this design allows the DNN to capture hierarchical and nonlinear correlations among vehicle variables, offering an effective complementary modeling method to SVR and CatBoost.

2.3.4. Hyperparameter tuning and overfitting mitigation

All three machine learning models were tuned utilizing a 5-fold cross-validation technique to guarantee consistent generalization performance throughout Bangkok's diverse vehicle fleet dataset, which consists of 1,846,111 records. Cross-validation was employed to reduce sample bias. It was also used to obtain performance estimates that are resilient across different vehicle categories, fuel types, and emission regulations.

For CatBoost, hyperparameter optimization focused on tree depth (6–10), learning rate (0.01–0.3), number of boosting iterations (500–2,000), and L2 leaf regularization (1–10). Random and grid searches were utilized according to the sensitivity of the parameters. The ordered boosting mechanism of CatBoost, which inhibits target leakage during categorical feature encoding, functioned as a principal method for mitigating overfitting, while early stopping based on validation loss was employed to avert superfluous iterations on the extensive dataset.

For SVR with RBF kernel, regularization parameter C (0.1–100), kernel coefficient γ (1e-4–1e-1), and ε-insensitive margin ε (0.001–0.1) were tuned via exhaustive grid-search. Because SVR exhibits high computational costs on datasets exceeding one million samples, representative stratified sub-sampling was applied to maintain feasibility while preserving the distribution of key vehicle attributes (fuel type, engine size, and Euro standard). The standardization of continuous predictors and the intrinsic margin-based loss function further aided in the regulation of overfitting.

The DNN's architecture was optimized by varying the number of hidden layers (2–5), neurons per layer (32–256), learning rate (0.0005–0.01), and dropout rates (0.1–0.4). The model employed ReLU activation with Adam optimization to facilitate efficient convergence in the high-dimensional feature space. Overfitting was alleviated with the implementation of dropout layers, L2 weight regularization, and early halting initiated by a plateau in validation performance. The batch size was determined by graphics processing unit (GPU) memory limitations to improve training efficiency and ensure gradient stability.

All hyperparameter ranges were derived from previous research on automobile emission modeling and refined through the first pilot experiments conducted on subsets of the Bangkok dataset. The tuning techniques were crucial for maintaining the models' generalizability across various vehicle categories. They also ensured robustness across realistic emission scenarios in metropolitan megacities.

3. EXPERIMENTAL

This section presents the experimental workflow used to construct, calibrate, and evaluate the machine learning models for estimating vehicle-level CO_2 emissions in Bangkok. The workflow consists of seven sequential phases: data acquisition, dataset integration, preprocessing, feature engineering, model development, hyperparameter optimization, and final evaluation, designed to ensure methodological transparency and full reproducibility. Figure 2 summarizes the complete analytical pipeline adopted in this study.

3.1. Data preparation and data integration

To develop a comprehensive vehicle-level database for modeling CO_2 emissions from the 1,846,111 vehicles registered in Bangkok, three complementary datasets were integrated: i) laboratory-tested fuel economy and CO_2 emission measurements provided by the U.S. EPA [11], ii) national vehicle registration records acquired from the DLT between January 2023 and January 2025, and [12], and iii) CO_2 emission factors obtained from the COPERT 5 model and the Euro 4-6 regulatory standards issued by the European Environment Agency [18]–[20], [24], [32]. The EPA dataset supplied standardized pollution benchmarks, engine specs, and fuel economy ratings. Meanwhile, the DLT dataset presented a detailed depiction of Bangkok's actual vehicle fleet makeup, encompassing brand, model, size class, engine configuration, and fuel type.

To merge these datasets, EPA and DLT entries were harmonized using shared identifiers such as manufacturer, model name, engine displacement, and fuel system attributes. Fuzzy-matching procedures were employed to resolve variations in naming conventions across sources and to identify the closest EPA equivalent for each registered vehicle. When a match to EPA data could not be established, COPERT-based CO_2 coefficients were assigned according to vehicle class, fuel type, and Euro certification level.

After integrating all sources, records lacking both EPA values and COPERT-assignable emission factors were removed. This filtering step produced 717,089 vehicles (38.8% of the original fleet) with complete and valid attributes required for CO_2 estimation and machine learning modeling. This integrated dataset formed a stable, internally consistent basis for subsequent analysis.

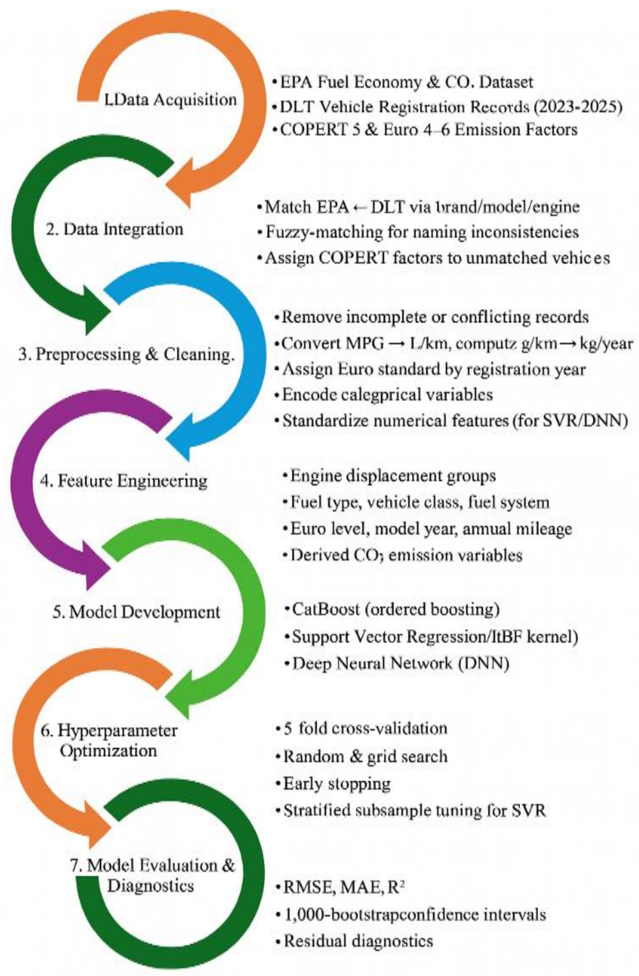

Figure 2. Workflow of CO_2 emissions estimation and machine learning analysis

3.2. Emissions estimation methodology

CO_2 emissions for each vehicle were estimated using internationally recognized methodologies, including the IPCC 2006 guidelines for National Greenhouse Gas Inventories [33], 2019 refinement [34], and the European Environment Agency's COPERT 5 framework [20] and the EPA dataset's fuel economy numbers were normalized by converting MPG to liters per kilometer (L/km) to guarantee consistency with emission factor estimates used in both the IPCC and COPERT approaches. For each vehicle record, annual

CO_2 emissions were computed as the product of fuel consumption, the corresponding emission factor (expressed in grams of CO_2 per liter of fuel), and an assumed annual driving distance of 15,000 km, consistent with regional transportation studies. The general estimation formula is expressed as (6).

$$CO_2 = Fuel\ Consumption\ \left(\frac{L}{km}\right) \times Emission\ Factor\ (g/L) \times 15000 \quad (6)$$

Where fuel consumption denotes the standardized fuel-use rate, and the emission factor represents the amount of CO_2 emitted per liter of fuel combusted. Emission factors vary according to fuel type, combustion characteristics, and Euro emission certification levels (Euro 4–6), following European Environment Agency COPERT framework and standardized greenhouse gas emission factor references [10], [20], [35]–[37]. This hybrid estimation approach, integrating empirical EPA-based emission benchmarks with COPERT-derived regulatory coefficients, ensures scientifically consistent and policy-relevant CO_2 estimates for all 717,089 registered vehicles, thereby providing a robust foundation for supervised machine learning analysis.

3.3. Data cleaning and preprocessing

A structured preprocessing pipeline was implemented to ensure that integrated EPA–DLT–COPERT dataset was consistent, complete, and analytically reliable. After merging the three sources, several inconsistencies and missing values were identified, particularly in fuel-economy measurements, CO_2 values, and fuel-type classifications. Furthermore, records lacking essential attributes required for emission estimation—such as fuel type, engine displacement, or EPA/COPERT emission values—were removed to avoid systematic bias. For variables with partial missingness (e.g., MPG for certain model-year groups), within-category median imputation was applied to preserve distributional characteristics.

Due to the absence of specific Euro emission criteria in DLT registrations, each vehicle was categorized into Euro 3–6 based on its registration year, adhering to legislative adoption timetables as shown in Table 3. This rule-based classification guaranteed alignment with COPERT emission factors. It also enabled uniform factor allocation across vehicle classifications.

To improve predictive performance, several engineered features were developed, including displacement categories, fuel-system classifications (ICE, hybrid, PHEV, and BEV), Euro certification levels, and derived emission metrics. Numerical variables were standardized for SVR and DNN to enhance convergence stability, whereas CatBoost utilized raw values owing to its resilience with varied tabular data. This multi-step preparation pipeline generated a refined, standardized dataset appropriate for extensive supervised learning, while ensuring methodological transparency and repeatability within Bangkok's diverse urban fleet.

Table 3. Euro emission standard

Registration year	Euro standard assigned
≤2009	Euro 3 or under
2010–2014	Euro 4
2015–2019	Euro 5
≥2020	Euro 6

3.4. Machine learning techniques

Three supervised machine learning algorithms, CatBoost, SVR, and DNN, were utilized to forecast vehicle-level CO_2 emissions from Bangkok's diverse urban fleet. These models were chosen for their complementary abilities to capture nonlinear patterns, manage mixed categorical and numerical structures, and simulate high-dimensional connections typically found in transportation-emission data. Previous studies have established the appropriateness of CatBoost for tabular datasets with high-cardinality categorical features [26], [27], the efficacy of DNN in capturing multiscale nonlinear interactions in environmental and engineering contexts [28]–[30], [38], and the theoretical resilience of SVR as a robust nonlinear benchmark for regression tasks [31].

CatBoost, a gradient-boosting decision-tree algorithm, was optimized with ordered boosting and target-based encoding to mitigate target leakage and diminish prediction bias. Its inherent capacity to process unrefined categorical variables, such as vehicle brand, model, fuel type, and Euro certification, negated the necessity for substantial encoding, rendering it particularly appropriate for Thailand's varied and high-cardinality vehicle register. Crucial hyperparameters, such as tree depth, learning rate, and the number of boosting iterations, were modified to improve generalization performance among fleet subgroups.

SVR with RBF kernel served as a theoretically grounded nonlinear baseline. The method is capable of modeling complex relationships among engine displacement, Euro standard, fuel type, and fuel-economy variables. Because SVR becomes computationally prohibitive on datasets exceeding one million samples, hyperparameter tuning was conducted on stratified subsamples that preserved the distributional characteristics of the overall fleet. Model parameters (C, γ, and ε) were optimized to balance flexibility and resilience against overfitting.

A fully connected feedforward DNN architecture was implemented to capture multilevel nonlinear relationships inherent in real-world emission behavior. The network employed ReLU activation, the Adam optimizer, and regularization techniques, including dropout, batch normalization, and early stopping, to stabilize training and control overfitting. Architectural parameters such as the number of hidden layers, units per layer, learning rate, and dropout ratios were tuned through cross-validation to ensure robust predictive performance.

All three models were trained utilizing the identically designed feature set delineated in sub-section 3.3 to guarantee equity in comparison assessment. The evaluation of the model performance utilized RMSE, MAE, and R^2. It facilitated a thorough and internally coherent comparison among ensemble-based, kernel-based, and deep learning methodologies.

3.5. Hyperparameter optimization

To improve forecast accuracy and guarantee strong generalization to Bangkok's diverse fleet, all three models underwent extensive hyperparameter optimization. Additionally, a 5-fold cross-validation method was used for all models to minimize sampling bias. It was crucial for generating accurate performance estimates given the dataset's diverse distributions of fuel types, engine displacements, and Euro standards.

CatBoost parameters were optimized using a combination of randomized search and grid refinement. Key parameters included tree depth (6–10), learning rate (0.01–0.3), number of boosting iterations (500–2,000), and L2 leaf regularization (1–10). Ordered boosting, which prevents target leakage when encoding high-cardinality categorical variables, served as an essential regularization mechanism. Early stopping was applied to avoid unnecessary boosting iterations once validation loss plateaued.

Optimization was performed on stratified subsamples that preserved the distribution of fuel type, engine displacement, and Euro standard, based on the computational expense of SVR on datasets. A thorough grid search optimized the penalty parameter C (0.1–100), kernel coefficient γ (1e-4–1e-1), and ε-insensitive margin ε (0.001–0.1). The standardization of continuous features ensured numerical stability, while the margin-based loss function was regularized to mitigate overfitting.

The hyperparameters of the DNN, network depth (2–5 layers), and number of neurons per layer (32–256), learning rate (0.0005–0.01), and dropout ratios (0.1–0.4), were optimized by iterative validation. The ReLU activation function and the Adam optimizer facilitated efficient convergence in high-dimensional feature spaces. Regularization methods, including dropout, L2 weight decay, and early stopping, were employed to enhance training stability and mitigate overfitting, which is particularly crucial due to the dataset's scale and diversity.

All hyperparameter configurations were initially assessed on subsets of the integrated database to ensure computational feasibility. Final tuning parameters were chosen based on their capacity to balance model accuracy. They also ensured convergence stability and generalization across vehicle categories, fuel systems, and emission levels.

3.6. Model evaluation and statistical testing

The predictive performance of the three machine learning models was evaluated using a comprehensive evaluation framework to ensure a rigorous and internally consistent comparison. The R^2, MAE, and RMSE, three widely utilized metrics for assessing regression quality in environmental and transportation modeling, were employed to quantify the models' accuracy. The metrics were computed under consistent experimental conditions to guarantee comparability among CatBoost, SVR, and DNN.

A nonparametric bootstrap approach with 1,000 resamples was employed for each statistic to address sampling variability within the extensive and diverse fleet dataset. The resultant 95% confidence intervals quantify model uncertainty and diminish reliance on a singular train-test division, thus enhancing the resilience and reliability of performance assessments. This methodology conforms to best practices in data-driven emission modeling, wherein assessing uncertainty is equally crucial as presenting point values.

In addition to the numerical results, the study also examined how each model behaved across different parts of the dataset. This was done through a set of residual-based diagnostics, including scatterplots, kernel-density curves, and error-distribution checks. These analyses were used to detect issues such as heteroscedasticity, persistent bias, or unusual error patterns in specific vehicle groups. Particular attention was given to segments that often behave differently, such as heavy-duty trucks, hybrid vehicles, and small-engine motorcycles, to confirm that no model performed well only because of certain dominant categories in the data.

By integrating conventional performance metrics with uncertainty analysis and detailed residual inspection, the evaluation offers a clearer and more grounded picture of how the models actually function. This layered approach provides a dependable basis for interpreting the results in later policy discussions. It also supports the application of the models to practical emission-forecasting scenarios.

4. RESULTS AND DISCUSSION

This section presents the empirical findings derived from the integrated EPA–DLT–COPERT dataset and the three machine learning models. Analyses encompass temporal fleet characteristics, fuel economy-emission correlations, feature associations, and model performance evaluated with uncertainty metrics. All outcomes directly correspond with the consolidated methodological framework depicted earlier in Figure 2.

4.1. Fleet characteristics and temporal patterns

Prior research on CO_2 emissions from vehicles has mostly used simulation-based models, like the COPERT model, or standardized laboratory tests to estimate emissions. However, because of unconsidered elements, including driving habits, traffic jams, and varying environmental conditions, such approaches frequently fall short in real-world scenarios. Furthermore, a lot of models don't incorporate real registration data or contemporary machine learning techniques that can deal with nonlinear and high-dimensional data. To overcome these constraints, this study combines laboratory-tested emissions, empirical registration data, and sophisticated machine learning techniques to increase the accuracy of CO_2 emission prediction in Bangkok, Thailand.

Figure 3 presents the temporal distribution of vehicle registrations across seven major categories and the violin plots visualize how each category is distributed over time, highlighting differences in registration concentration rather than absolute counts. Private passenger cars with not more than seven seats exhibit the widest and most concentrated density, indicating that this category consistently accounts for a substantial share of newly registered vehicles. Private passenger vehicles with more than seven seats and private freight vehicles show moderately wide distributions, suggesting steady but less dominant registration patterns. Categories related to public transit, such as private buses and taxis, demonstrate relatively narrow and stable distributions, reflecting their steady operational demand. Rented three-wheeled vehicles demonstrate a concentrated density in 2023, but commercial service vehicles show a broader distribution extending into late 2024 and early 2025, reflecting a more diverse registration timetable within this category. The temporal distribution reveals a fleet composition predominantly consisting of light-duty private vehicles, but commercial and public transport categories display a lesser, still consistent presence. The discrepancies in category-specific registration trends are substantial, as they directly influence the overall CO_2 emission profile of Bangkok's vehicle fleet.

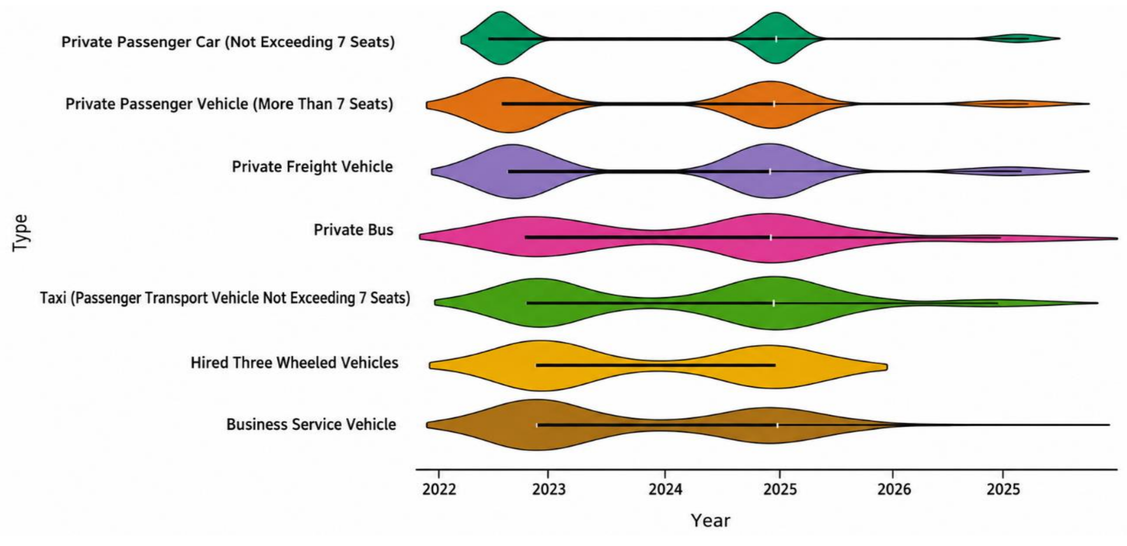

Figure 3. Temporal distribution of vehicle types in Thailand

4.2. Relationships between fuel consumption and CO_2 emissions

Figure 4 illustrates the relationship between combined fuel consumption (L/100 km) and CO_2 emissions (g/km) across three major fuel types: gasoline, diesel, and hybrid vehicles. A strong positive association is apparent. Vehicles with higher fuel consumption per distance consistently generate higher CO_2 emissions, reflecting the fundamental carbon content of the fuels.

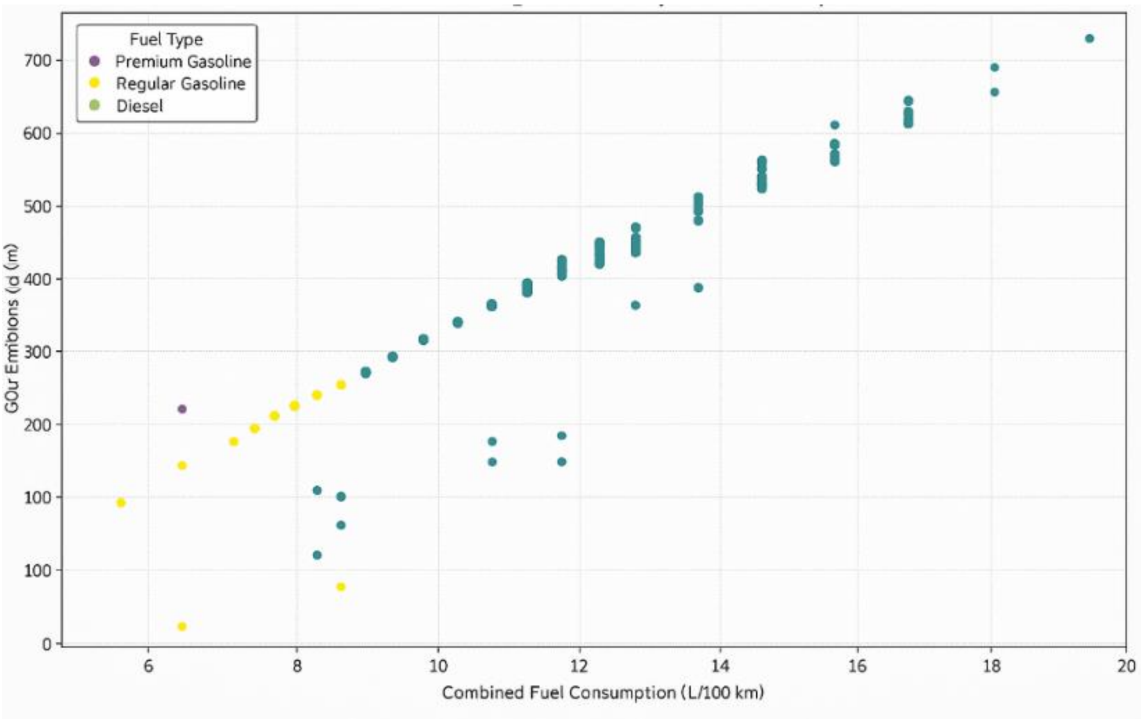

Figure 4. CO_2 emissions by fuel consumption

Gasoline vehicles cluster toward the higher end of both fuel consumption and CO_2 output, indicating less efficient combustion relative to other categories. Diesel vehicles exhibit a wider dispersion: several models achieve comparatively lower CO_2 emissions due to higher thermal efficiency, while others display emission levels similar to gasoline vehicles, which may be attributed to greater engine loads, aging technologies, or deteriorated emission-control systems. Hybrid vehicles occupy the lower region of the distribution, confirming reduced tailpipe emissions resulting from electric-motor assistance.

The benchmark CO_2 ranges by fuel type in Table 4, which are derived from EPA, COPERT, and regional studies, provide context for these trends. With gasoline vehicles typically falling between 215 and 240 g/km, diesel vehicles between 260 and 320 g/km, hybrids between 95 and 130 g/km, motorbikes between 65 and 90 g/km, and heavy-duty vehicles between 380 and 480 g/km, the observed emission levels in the Bangkok fleet nearly match these international reference values. These comparisons substantiate the credibility of the combined EPA–COPERT estimating approach and underscore the impact of fuel characteristics, combustion efficiency, and vehicle technology on actual emission performance. These relationships confirm that fuel type and driving efficiency remain primary determinants of CO_2 performance across the Bangkok fleet.

Table 4. Typical CO_2 emission ranges by fuel type

Fuel type	Mean CO_2 (g/km)	Typical range from literature	Interpretation
Gasoline	215–240	EPA: 210–260	Consistent with light-duty vehicle norms
Diesel	260–320	COPERT: 250–350	Higher emissions due to fuel density
Hybrid	95–130	EPA: 80–120	Reduced engine load
Motorcycle	65–90	ADB/Thailand 2022	Small displacement
Heavy-duty	380–480	COPERT: 350–500	Large variation from load conditions

4.3. Feature importance and correlation structure

The most significant predictors of vehicle-level CO_2 emissions were identified using RFE, which improves model interpretability and reduces redundancy. RFE was applied using a random forest regressor to iteratively rank and eliminate the least informative variables due to the high dimensionality of the integrated EPA–DLT–COPERT dataset. This dataset includes fuel consumption metrics, tailpipe CO_2 values, fuel economy indicators, fuel types, and emission factors.

Throughout multiple cycles, RFE consistently identified variables directly linked to combustion intensity and energy consumption as the most significant predictors. This encompasses annual petroleum consumption, tailpipe CO_2 emissions from EPA data, combined/city/highway MPG, and the COPERT-derived EF_CO_2 factor. These picks accord with the physical laws regulating carbon emissions, wherein increased fuel use and diminished fuel efficiency directly relate to heightened emissions.

Figure 5 illustrates these correlations through a correlation heatmap. Fuel-economy measurements (city, highway, and combined MPG) have substantial negative correlations with CO_2 emissions (r =–0.77 to –0.82), demonstrating the expected inverse link between fuel efficiency and carbon emissions. In contrast, annual fuel consumption and EPA tailpipe CO_2 have a strong positive correlation (r =0.49) and the Euro-adjusted emission factor EF_CO_2 demonstrates a moderate positive correlation, affirming its importance in regulatory evaluation. Taken together, the RFE outcomes and correlation structure confirm that the engineered dataset captures key mechanistic drivers of CO_2 emissions. They also validate the feature selection and preprocessing technique detailed in section 3, ensuring that subsequent machine learning models depend on predictors that are both statistically sound and scientifically significant.

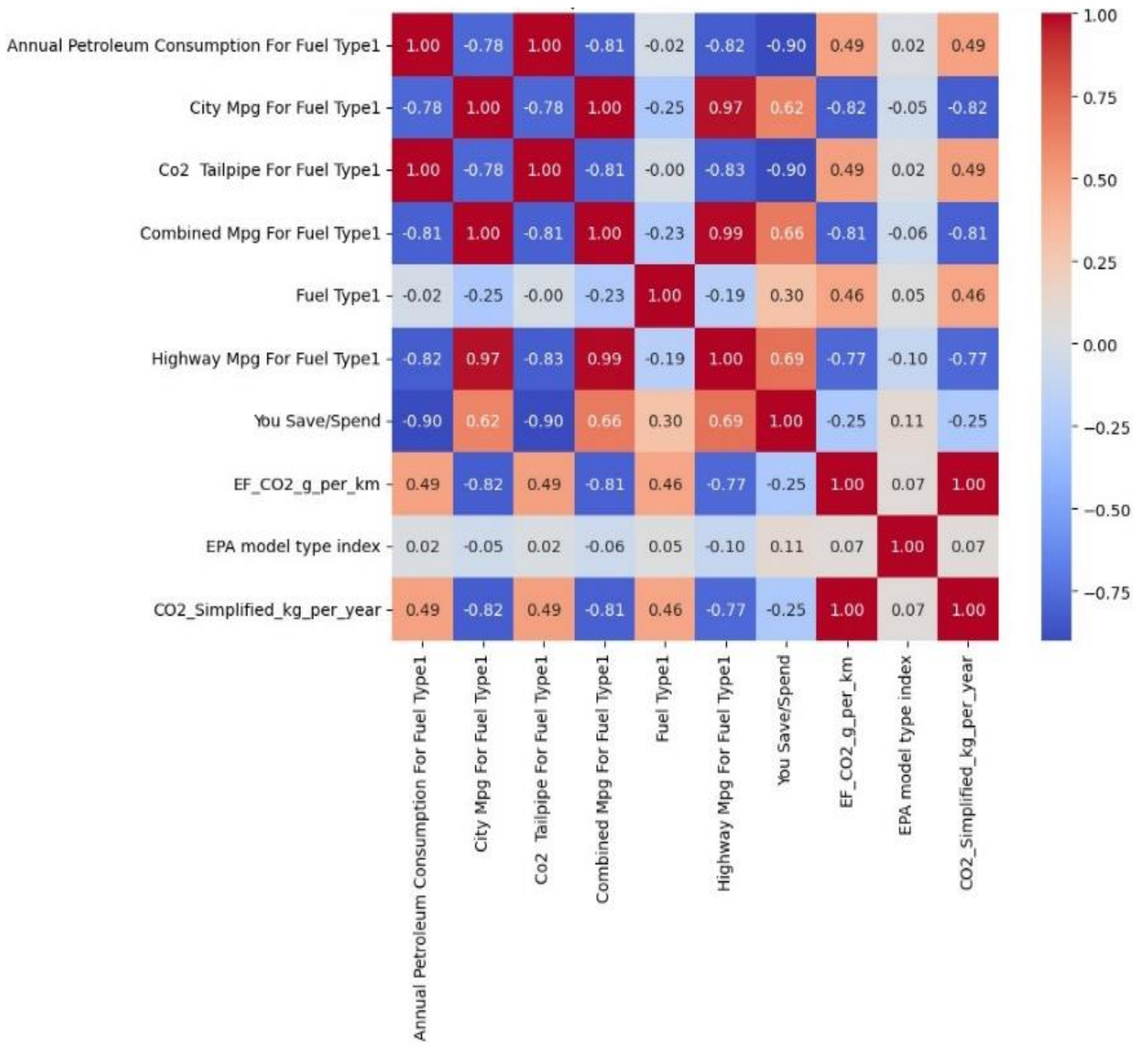

Figure 5. Correlation heatmap of selected features and CO_2 emissions

4.4. Model performance comparison

Table 5 displays the predictive accuracy of the three machine learning models—CatBoost, DNN, and SVR, utilizing the integrated EPA–DLT–COPERT dataset. CatBoost demonstrates superior predictive accuracy, attaining the lowest RMSE and MAE, alongside the greatest R^2, indicative of its proficiency in managing heterogeneous tabular data and high-cardinality categorical variables, and the DNN model exhibits comparable performance, albeit with significantly reduced accuracy, particularly in tail-distribution scenarios where deep neural architectures may be susceptible to infrequent vehicle classifications. Also, SVR exhibits the poorest performance, aligning with its established constraints on extensive nonlinear datasets. The findings indicate that ensemble learning and deep learning methodologies are superior to kernel-based techniques for extensive, diverse transportation datasets.

Table 5. Model performance metrics

Model	RMSE (g/km)	95% CI	MAE (g/km)	R^2
CatBoost	23.4	22.7–24.5	14.6	0.921
DNN	28.1	27.0–29.6	18.2	0.887
SVR (RBF)	32.8	31.4–34.3	21.0	0.846

Figure 6 presents a comparison of actual CO_2 emissions and projected values for a sample of 2025 automobiles. The proximity of the observed and predicted bars for the majority of vehicle entries implies a robust fit and dependable generalization, and slight disparities likely emerge from unobserved variables (e.g., aftermarket modifications and actual driving loads) not included in the current datasets. This convergence between actual and expected values demonstrates the appropriateness of the proposed modeling framework for practical emission estimation and fleet-level forecasting.

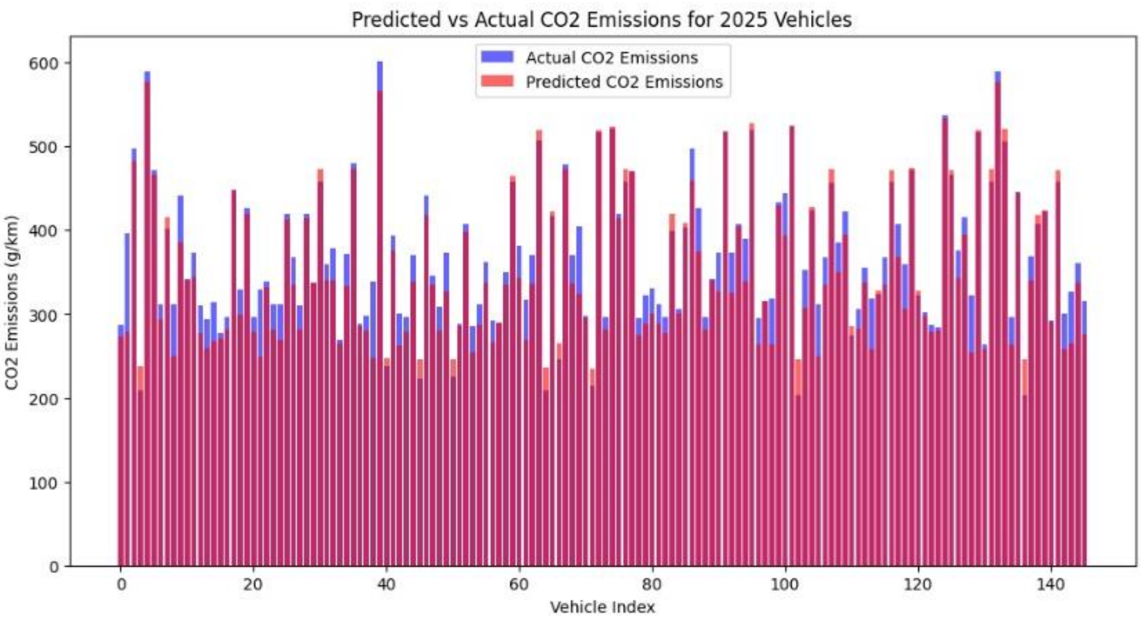

Figure 6. Comparison of actual and predicted CO_2 emissions (g/km) for 2025 vehicle registrations

4.6. Implications for transport and environmental policy

The prevalence of lightweight private vehicles indicates that policies focused on fuel efficiency, engine technology, and electricity in this sector would produce the most significant emission reductions and the persistent increase in diesel-fueled business vehicles underscores potential for focused retrofitting or expedited transition to greener options. In addition, the robust effectiveness of data-driven models like CatBoost and DNN illustrates that machine learning methodologies can enhance conventional inventory techniques such as COPERT by identifying nonlinearities and real-fleet attributes that traditional models could neglect. Furthermore, incorporating predictive methods into national monitoring systems could facilitate more dynamic, evidence-based policy actions, such as pinpointing high-emission fleet segments or directing incentive programs for lower-emission technology.

5. CONCLUSION

This research illustrates the significance of integrating real-world vehicle registration data with laboratory-derived emission metrics and machine learning methodologies to assess CO_2 emissions at the individual car level for Bangkok's diverse fleet. Through integrating EPA fuel-economy records, COPERT/Euro emission factors, and over 1.8 million DLT registrations, the resultant dataset encapsulated essential technical characteristics, fuel consumption, emission factors, displacement, fuel type, and Euro standards, recognized for influencing real-world emission behavior. Also, the implementation of RFE enhanced model interpretability by consistently identifying fuel-consumption variables, tailpipe CO_2, and emission-factor characteristics as the most significant predictors, in accordance with the fundamental principles of fuel combustion. CatBoost surpassed DNN (R^2 =0.887) and SVR (R^2 =0.846), demonstrating superior prediction accuracy (RMSE =23.4 g/km, R^2 =0.921). The enhanced accuracy and stability observed were attributed to CatBoost's proficiency in handling heterogeneous tabular data, nonlinear relationships, and categorical variables characterized by high cardinality. Although DNN can capture complex multiscale patterns, its performance was somewhat inferior to that of CatBoost and SVR, which exhibited elevated error rates and reduced scalability even after hyperparameter adjustments, highlighting its limitations for large and diverse vehicle datasets.

6. FUTURE WORK

Future extensions of this work should incorporate more granular temporal and spatial data, such as daily driving profiles, congestion levels, and road-type classifications, to enhance the generalizability of the models. Additional contextual variables, including meteorological conditions, traffic intensity, and vehicle-usage patterns, may further strengthen predictive accuracy and enable scenario-based forecasting. Employing explainable artificial intelligence (XAI) approaches such as Shapley additive explanations (SHAP) could provide clearer insights into model reasoning, supporting transparent and evidence-based policymaking. Finally, integrating these models into operational platforms, like transport-planning dashboards, emissions-monitoring systems, or real-time mobility applications, would offer practical tools for regulatory agencies and urban planners seeking to reduce transportation-related emissions and support Thailand's long-term climate commitments.

FUNDING INFORMATION

This research was funded by Suan Sunandha Rajabhat University.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Phachaya Chaiwchan		✓				✓		✓	✓	✓	✓	✓		
Kunyanuth Kularbphettong	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are publicly available from the following source:

- https://public.opendatasoft.com/explore/dataset/all-vehicles-model/.
- https://datagov.mot.go.th/dataset/dataset_stat_1_001/resource/3797ba01-cbc5-4de1-ac3b-de9fe37887b0.

REFERENCES

[1] Thailand's, *Thailand's long-term low greenhouse gas emission development strategy (Revised version)*, Bangkok, Thailand: Office of Natural Resources and Environmental Policy and Planning; and Ministry of Natural Resources and Environment, 2022.

[2] S. Nilrit, P. Sampanpanish, and S. Bualert, "Carbon dioxide and methane emission rates from taxi vehicles in Thailand," *Carbon Management*, vol. 9, no. 1, pp. 37–43, Jan. 2018, doi: 10.1080/17583004.2017.1412232.

[3] S. Nilrit, P. Sampanpanish, and S. Bualert, "Comparison of CO2 emissions from vehicles in Thailand," *Applied Environmental Research*, vol. 39, no. 1, pp. 65–74, Apr. 2017.

[4] G. F. Giuzio, C. Forzano, G. Barone, and A. Buonomano, "Accelerating the low-carbon transition: technological advancements and challenges for the sustainable development of energy, water, and environment systems," *Energy Reports*, vol. 11, pp. 4676–4687, Jun. 2024, doi: 10.1016/j.egyr.2024.04.031.

[5] World Health Organization, "Ambient (outdoor) air pollution," *WHO.int*, 2024. Accessed: Jan. 31, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[6] Y. Laoonual and K. Nantasaksiri, "The green revolution in Thailand steering by electric buses," *slocat.net*, 2023. Accessed: Jan. 31, 2025. [Online]. Available: https://slocat.net/the-green-revolution-in-thailand-steering-by-electric-buses/

[7] S. Ongkittikul, "Braving hurdles for green transport," *tdri.or.th*, 2024. Accessed: Jan. 31, 2025. [Online]. Available: https://tdri.or.th/en/2024/01/braving-hurdles-for-green-transport/

[8] D. Lejri, A. Can, N. Schiper, and L. Leclercq, "Accounting for traffic speed dynamics when calculating COPERT and PHEM pollutant emissions at the urban scale," *Transportation Research Part D: Transport and Environment*, vol. 63, pp. 588–603, Aug. 2018, doi: 10.1016/j.trd.2018.06.023.

[9] S. Dey, B. Caulfield, and B. Ghosh, "Modelling uncertainty of vehicular emissions inventory: a case study of Ireland," *Journal of Cleaner Production*, vol. 213, pp. 1115–1126, Mar. 2019, doi: 10.1016/j.jclepro.2018.12.125.

[10] O. Orynycz, M. Z. -Laskowska, and E. Kulesza, "CO_2 emission and energy consumption estimates in the COPERT model—conclusions from chassis dynamometer tests and SANN artificial neural network models and their meaning for transport management," *Energies*, vol. 18, no. 13, Jul. 2025, doi: 10.3390/en18133457.

[11] D. Sirithian, P. Thanatrakolsri, and S. Pongpan, "CO_2 and CH_4 emission factors from light-duty vehicles by fuel types in Thailand," *Atmosphere*, vol. 13, no. 10, Sep. 2022, doi: 10.3390/atmos13101588.

[12] M. Kuson, J. Mahujchariyawong, T. Satapanajaru, and T. Prueksasit, "Assessment and management of air pollutant emissions from vehicles in the Bangkok metropolitan region," *Science & Technology Asia*, vol. 28, no. 4, pp. 144–155, 2023.

[13] D. Yuan, L. Tang, X. Yang, F. Xu, and K. Liu, "Explainable machine learning prediction of vehicle CO_2 emissions for sustainable energy and transport," *Energies*, vol. 18, no. 20, Oct. 2025, doi: 10.3390/en18205408.

[14] F. Gurcan, "Forecasting CO_2 emissions of fuel vehicles for an ecological world using ensemble learning, machine learning, and deep learning models," *PeerJ Computer Science*, vol. 10, Aug. 2024, doi: 10.7717/peerj-cs.2234.

[15] G. Çınarer, M. K. Yeşilyurt, Ü. Ağbulut, Z. Yılbaşı, and K. Kılıç, "Application of various machine learning algorithms in view of predicting the CO_2 emissions in the transportation sector," *Science and Technology for Energy Transition*, vol. 79, Mar. 2024, doi: 10.2516/stet/2024014.

[16] G. Zhou, L. Mao, T. Bao, and F. Zhuang, "Machine learning-driven CO2 emission forecasting for light-duty vehicles in China," *Transportation Research Part D: Transport and Environment*, vol. 137, Dec. 2024, doi: 10.1016/j.trd.2024.104502.

[17] S. E. Yalçın, "Estimation of CO_2 emissions in transportation systems using artificial neural networks, machine learning, and deep learning: a comprehensive approach," *Systems*, vol. 13, no. 3, Mar. 2025, doi: 10.3390/systems13030194.

[18] European Commission, "Computer model to calculate emissions from road traffic," *web.jrc.ec.europa.eu*, 2025. Accessed: Jan. 29, 2025. [Online]. Available: https://web.jrc.ec.europa.eu/policy-model-inventory/explore/models/model-copert/

[19] N. Kholod *et al.*, "A methodology for calculating transport emissions in cities with limited traffic data: case study of diesel particulates and black carbon emissions in Murmansk," *Science of The Total Environment*, vol. 547, pp. 305–313, Mar. 2016, doi: 10.1016/j.scitotenv.2015.12.151.

[20] L. Ntziachristos, D. Gkatzoflias, C. Kouridis, and Z. Samaras, "COPERT: a European road transport emission inventory model," in *Information Technologies in Environmental Engineering*, Berlin, Heidelberg: Springer, 2009, pp. 491–504, doi: 10.1007/978-3-540-88351-7_37.

[21] Y. Natarajan, G. Wadhwa, K. R. S. Preethaa, and A. Paul, "Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms," *Electronics*, vol. 12, no. 10, May 2023, doi: 10.3390/electronics12102288.

[22] Environmental Protection Agency, "All vehicles model," *public.opendatasoft.com*, 2024. Accessed: Jan. 12, 2025. [Online]. Available: https://public.opendatasoft.com/explore/assets/all-vehicles-model/

[23] MOT Data Catalog, "Statistics on the number of first-time registered vehicles according to the law on motor vehicles-classification by brand and model (monthly) August 2024," *datagov.mot.go.th*, 2024. Accessed: Jan. 12, 2025. [Online]. Available: https://datagov.mot.go.th/dataset/dataset_stat_1_001/resource/3797ba01-cbc5-4de1-ac3b-de9fe37887b0

[24] European Commission, "Emissions in the automotive sector," *single-market-economy.ec.europa.eu*, 2022. Accessed: Dec. 31, 2024. [Online]. Available: https://single-market-economy.ec.europa.eu/sectors/automotive-industry/environmental-protection/emissions-automotive-sector_en

[25] G. Fontaras, N.-G. Zacharof, and B. Ciuffo, "Fuel consumption and CO_2 emissions from passenger cars in Europe – Laboratory versus real-world emissions," *Progress in Energy and Combustion Science*, vol. 60, pp. 97–131, 2017, doi: 10.1016/j.pecs.2016.12.004.

[26] H. Sun, J. He, Y. Chen, and B. Zhao, "Space-time sea surface pCO_2 estimation in the North Atlantic based on CatBoost," *Remote Sensing*, vol. 13, no. 14, Jul. 2021, doi: 10.3390/rs13142805.

[27] X. Feng, J. He, and B. Lu, "Accurate and generalizable soil liquefaction prediction model based on the CatBoost algorithm," *Acta Geophysica*, vol. 72, no. 5, pp. 3417–3426, Jun. 2024, doi: 10.1007/s11600-024-01381-9.

[28] A. B. Ghorbal, A. Grine, I. Elbatal, E. M. Almetwally, M. M. Eid, and E.-S. M. El-Kenawy, "Predicting carbon dioxide emissions using deep learning and Ninja metaheuristic optimization algorithm," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-86251-0.

[29] E. Komara, Z. Tsaniyah, and W. Utama, “Prediction of CO2 gas saturation distribution based on deep learning using deep neural network (DNN) algorithm,” *Jurnal Geosaintek*, vol. 9, no. 2, Aug. 2023.

[30] D. T. -Gago, G. Golcarenarenji, I. M. -Alpiste, Q. Wang, and J. M. A. -Calero, “Machine-learning-based carbon dioxide concentration prediction for hybrid vehicles,” *Sensors*, vol. 23, no. 3, Jan. 2023, doi: 10.3390/s23031350.

[31] A. H. Al-Nefaie and T. H. H. Aldhyani, “Predicting CO_2 emissions from traffic vehicles for sustainable and smart environment using a deep learning model,” *Sustainability*, vol. 15, no. 9, May 2023, doi: 10.3390/su15097615.

[32] M. Ali, M. D. Kamal, A. Tahir, and S. Atif, “Fuel consumption monitoring through COPERT model—a case study for urban sustainability,” *Sustainability*, vol. 13, no. 21, Oct. 2021, doi: 10.3390/su132111614.

[33] Intergovernmental Panel on Climate Change, “2006 IPCC guidelines for National Greenhouse Gas Inventories-volume 2: energy,” *ipcc-nggip.iges.or.jp*, 2006. Accessed: Dec, 31. 2024. [Online]. Available: https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol2.html

[34] Intergovernmental Panel on Climate Change, “2019 refinement to the 2006 IPCC guidelines for National Greenhouse Gas Inventories,” *ipcc-nggip.iges.or.jp*, 2019. Accessed: Dec, 31. 2024. [Online]. Available: https://www.ipcc-nggip.iges.or.jp/public/2019rf/

[35] C. Samaras, L. Ntziachristos, and Z. Samaras, “COPERT Micro: a tool to calculate vehicle emissions in urban areas,” in *Energy and Environment*, Wiley, 2016, pp. 401–415, doi: 10.1002/9781119307761.ch26.

[36] U.S. Environmental Protection Agency, “GHG emission factors hub,” *epa.gov*, 2026. Accessed: Dec, 31. 2024. [Online]. Available: https://www.epa.gov/climateleadership/ghg-emission-factors-hub

[37] S. Kawsar, S. Biswas, M. Noor, and M. S. Mamun, “Investigating the applicability of COPERT 5.5 emission software in Bangladesh and developing countrywide vehicular emission inventories,” *Environmental Science: Atmospheres*, vol. 4, no. 1, pp. 57–72, 2024, doi: 10.1039/d3ea00047h.

[38] K. Bussaban, K. Kularbphettong, and C. Boonseng, “Prediction of CO2 emissions using machine learning,” *CONECT. International Scientific Conference of Environmental and Climate Technologies*, May 2023, doi: 10.7250/CONECT.2023.099.

BIOGRAPHIES OF AUTHORS

Phachaya Chaiwchan is a lecturer of Digital Innovation Management Program, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand. She is a Ph.D. candidate in Information Science, School of Liberal Arts, Sukhothai Thammathirat Open University (STOU). Her research interests include artificial intelligence, digital ethics, knowledge extraction, and decision making. She can be contacted at email: phachaya.ch@ssru.ac.th.

Kunyanuth Kularbphettong is an associate professor of Computer Science at Suan Sunandha Rajabhat University, Bangkok, Thailand. Indonesia. She received her Ph.D. in Information Technology from King Mongkut’s University of Technology North Bangkok (KMUTNB). Her research interests are environmental management, machine learning, deep learning, multi-agent, and educational mining. She can be contacted at email: kunyanuth.ku@ssru.ac.th.