International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

An India-Specific Hybrid Machine Learning Framework for Real-Time CO_2 Emission Prediction and Policy-Aware Analysis of Motor Vehicles

N. Shanmukha Priya[1], S. Jayasri[1], P. Siri Sai Sneha[1], K. Likitha[1]

[1]Department of CSE (AI&ML), Gayatri Vidya Parishad College of Engineering for Women(A), Kommadi, Madhurawada, Visakhapatnam – 530048

Abstract

Transportation contributes significantly to global greenhouse gas emissions, with CO_2 from motor vehicles being a dominant factor. This study proposes an India-specific machine learning framework for predicting vehicular CO_2 emissions using key operational and environmental features, including engine characteristics, fuel consumption, traffic conditions, and ambient factors. A structured dataset incorporating both vehicle parameters and Indian Driving Cycle (IDC) corrections was used for model development; the dataset is simulated/structured and does not yet incorporate large-scale real-world OBD-II sensor streams.

Multiple models — Linear Regression, Random Forest, Gradient Boosting, ARIMA, and an LSTM-inspired approximation model — were implemented and evaluated using MAE, RMSE, and R^2 metrics. Among them, Random Forest achieved the best performance ($R^2 \approx 0.873$), though results are interpreted within a controlled dataset setting and should be validated on real-world data before deployment. Temporal models such as ARIMA and the LSTM-inspired estimator capture short-term emission dynamics effectively.

The framework **integrates (in a prototype implementation)** environmental corrections and regulatory checks (BS6 approximate proxy compliance), enabling policy-aware predictions. An ablation study confirms that IDC correction factors and traffic modelling contribute the largest accuracy gains. SHAP (SHapley Additive exPlanations) analysis identifies mass air flow and engine speed as the dominant predictors. While results demonstrate strong predictive capability within the structured dataset, further validation on real-world OBD-II datasets is required. The proposed prototype framework provides a scalable foundation for intelligent emission monitoring and sustainable transportation planning in India.

Keywords: CO_2 Emissions, Machine Learning, Vehicle Emissions, Environmental Sustainability, Regression Models, Ensemble Learning, Random Forest, Gradient Boosting, Time Series Forecasting, LSTM-Inspired Approximation, Feature Engineering, SHAP, Hyperparameter Tuning, Indian Driving Cycle.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

1. Introduction

The rapid expansion of the transportation sector has significantly increased greenhouse gas emissions, making it one of the major contributors to global climate change. Among these emissions, carbon dioxide (CO_2) plays a dominant role, especially from motor vehicles. Monitoring and predicting these emissions have become essential for environmental protection, sustainable development, and effective policy implementation.

Conventional emission estimation methods are typically based on predefined analytical formulas and controlled laboratory testing. While these approaches provide baseline values, they fail to capture real-world variations such as stop-and-go driving, varying vehicle performance, and diverse fuel efficiencies. Machine learning (ML) techniques **have the potential to overcome** these limitations by learning complex relationships from historical data to generate more accurate, real-world-oriented predictions.

1.1 Motivation

India is experiencing a rapid surge in vehicle ownership, with 334.8 million registered vehicles as of 2024 — a 136% increase since 2010 (MoRTH, 2024). Road transport contributes 245.8 Mt of CO_2 annually, representing 86.8% of total transport-sector emissions. Urban air quality has reached critical levels, as shown in Table 1.

Table 1: Air Quality Index and $PM_{2.5}$ Levels in Selected Indian Cities (2023)

City	AQI	$PM_{2.5}$ (µg/m³)	WHO Safe Limit	Exceedance
Delhi	215	95.8	15	6.4×
Kanpur	210	90.6	15	6.0×
Agra	208	88.4	15	5.9×
Lucknow	201	83.2	15	5.5×
Varanasi	198	78.6	15	5.2×

Source: CPCB (2023); WHO Safe Limit per WHO (2021)

Despite India's NDC commitment to reduce emission intensity by 45% by 2030, the limited availability of validated India-specific real-time CO_2 prediction systems adapted for Indian roads, vehicles, and the Indian Driving Cycle (IDC) motivates this work.

1.2 Problem Definition

The core problem is the limited availability of a validated India-specific, real-time CO_2 emission prediction system. Key challenges include:

1. Lab-to-road gap: WLTP test cycle ratings understate real-world Indian emissions by 25–52% due to stop-and-go traffic, with idle times approaching 30%.
2. Fleet mismatch: Existing tools model cars only; India's fleet is dominated by two-wheelers (80.9% of sales) and CNG three-wheelers.

3. Regulatory gap: No real-time BS6 compliance checker exists despite BS6 being mandatory since April 2020.
4. Environmental blindness: Ambient temperature (routinely 35–45°C), humidity, and AQI affect engine combustion but are ignored by static tools.
5. Currency: Most existing tools report costs in USD/CAD, not INR.

The emission gap between test and real-world conditions is quantified as an empirical approximation based on IDC correction factors — similar correction approaches are widely used in transport emission studies (e.g., Tena-Gago et al., 2023):

$$\Delta CO_2 = CO_{2,WLTP} \times (f_{IDC} - 1) \quad (1)$$

For Bangalore ($f_{IDC} = 1.52$):

$$\Delta CO_2 = 120 \times (1.52 - 1) = 62.4 \text{ g/km (a 52\%understatement)} \quad (2)$$

2. Literature Survey

Research on vehicle CO_2 emission prediction has evolved from simple regression models to sophisticated deep learning architectures. Classical regression approaches (Sharma & Kumar, 2022) established that features such as engine displacement, fuel type, and fuel consumption can be linked to emissions via supervised learning, though these studies lack real-time adaptability. Ensemble methods (Al-Nefaie & Aldhyani, 2023; Kumari & Singh, 2023) demonstrated that Random Forest, Gradient Boosting Machines (GBM), and Support Vector Regression (SVR) significantly outperform linear baselines by capturing nonlinear feature interactions, at the cost of increased computational overhead. Deep learning temporal models (Alam et al., 2025; Li et al., 2024) proposed CNN+LSTM+Attention hybrids and GNN+LSTM architectures that model both spatial road-network dependencies and time-varying emission patterns. The reference LSTM architecture (Tena-Gago et al., *Sensors* 2023) trained on Toyota Prius OBD-II data achieves $R^2 = 0.975$ with a 32-step lookback window. IoT-integrated frameworks (Udoh et al., 2024) combine Decision Trees and KNN with live sensor streams for adaptive real-time inference, but reliability depends heavily on sensor quality.

Table 2: Literature Survey Summary

Study	Venue	Approach	Outcome	Limitation
Tena-Gago et al. (2023)	*Sensors* (MDPI)	UWS-LSTM	R^2=0.975, real OBD-II	Single vehicle type
Al-Nefaie & Aldhyani (2023)	*Appl. Sci.*	RF, GBM, SVR	RF best performer	Lab data only
Alam et al. (2025)	*IEEE Access*	CNN+LSTM+Attention	Higher accuracy than standalone LSTM	High complexity
Li et al. (2024)	*Trans. Res. Part*	GNN+LSTM	Outperforms ARIMA & LSTM	Complex graph preprocessing

Study	Venue	Approach	Outcome	Limitation
	C			
Udoh et al. (2024)	*IEEE IoT J.*	DT, KNN + IoT	Better real-time accuracy	Sensor noise sensitivity

3. Proposed System

The proposed system is a **prototype hybrid, modular CO_2 prediction framework** integrating five models into a unified prediction engine.

3.1 System Architecture

The architecture **integrates (in a prototype implementation)** multiple external data sources — WAQI API (air quality), Open-Meteo (weather), OBD-II adapters (vehicle telemetry), and PPAC/BEE (regulatory data) — into a centralised prediction engine. In the current implementation, these sources are represented by structured/simulated dataset entries rather than live API connections. The engine applies the appropriate ML or time-series model and returns CO_2 emission estimates (g/km), BS6 approximate proxy compliance status, and eco-friendly grades.

3.2 Feature Engineering

The 11-dimensional input feature vector $\mathbf{x} \in \mathbb{R}^{11}$, **constructed from a simulated/structured dataset**, is defined as:

$$\mathbf{x} = \left[v,\ \omega,\ \dot{m}_{\text{air}},\ \dot{m}_{\text{exhaust}},\ T_{\text{coolant}},\ M_{\text{engine}},\ a,\ \text{SOC},\ \text{AQI}_{\text{amb}},\ f_{\text{trac}},\ T_{\text{ambient}}\right] \quad (3)$$

3.2.1 Z-Score Normalisation

All features are standardised prior to model training:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (4)$$

where μ_i and σ_i are the mean and standard deviation of feature i computed over the training set.

3.2.2 IDC Temperature Correction Factor

Ambient temperature affects volumetric efficiency and fuel-air mixture. The empirical correction factor proposed in this study is:

$$f_{\text{temp}} = 1 + \max(0,\ (T_{\text{ambient}} - 25) \times 0.004) \quad (5)$$

At the typical Indian summer temperature of $T_{\text{ambient}} = 35°\text{C}$:

$$f_{\text{temp}} = 1 + (35 - 25) \times 0.004 = 1.04 \Longrightarrow 4\%\ CO_2 \text{ increase} \quad (6)$$

3.2.3 Traffic Congestion Time Model

The traffic factor varies by city c and hour-of-day h. The scaling multipliers (1.20 for peak hours, 1.05 for lunch hours) are values chosen based on typical congestion patterns reported in Indian urban traffic studies (MoRTH, 2024; CPCB, 2023):

$$f_{\text{trac}}(c,h) = \begin{cases} f_{\text{base}}(c) \times 1.20 & \text{if } h \in [8,10] \cup [17,20] \text{ (peak hours)} \\ f_{\text{base}}(c) \times 1.05 & \text{if } h \in [12,14] \text{ (lunch hours)} \\ f_{\text{base}}(c) & \text{otherwise (off-peak)} \end{cases} \quad (7)$$

4. Model Implementations

4.1 Linear Regression (Baseline)

Linear Regression models CO_2 emission as a linear combination of normalised input features:

$$\hat{y} = \mathbf{w}^{\top}\mathbf{z} + b = \sum_{i=1}^{11} w_i z_i + b \quad (8)$$

where $\mathbf{w} \in \mathbb{R}^{11}$ is the learned weight vector and $b \in \mathbb{R}$ is the bias term. Coefficients are estimated by minimising the mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \quad (9)$$

Limitations noted: The linear formulation does not include regularisation (Ridge/Lasso) and does not address multicollinearity between correlated OBD-II features (e.g., speed and RPM).

4.2 Random Forest Regressor

Random Forest builds an ensemble of T decision trees $\{h_1(\mathbf{z}), h_2(\mathbf{z}), \ldots, h_T(\mathbf{z})\}$, each trained on a bootstrap sample $\mathcal{D}_t \sim \mathcal{D}$ (with replacement). The final prediction is obtained by averaging:

$$\hat{y}_{\text{RF}}(\mathbf{z}) = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{z}) \quad (10)$$

4.2.1 Node Splitting

At each node m, a random subset of $k = \lfloor\sqrt{11}\rfloor = 3$ features is selected. The optimal split (j^*, θ^*) minimises the variance reduction (appropriate for regression tasks):

$$(j^*, \theta^*) = \underset{j,\theta}{\operatorname{argmin}} \frac{|S_L|}{|S|}\operatorname{Var}(S_L) + \frac{|S_R|}{|S|}\operatorname{Var}(S_R) \quad (11)$$

where S_L and S_R are the left and right child node samples, respectively.

4.2.2 Feature Importance (Variance-Reduction-Based)

The importance of feature j in tree t is accumulated over all nodes m where feature j is used. For **regression**, this is based on **variance reduction** (not Gini impurity, which applies to classification):

$$\text{Imp}_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{m:\,\text{split on } j} \Delta I_m^{(t)} \tag{12}$$

where $\Delta I_m^{(t)}$ is the weighted variance reduction at node m in tree t.

4.2.3 Bias–Variance Decomposition

Random Forest reduces prediction variance through tree averaging while keeping bias low:

$$\mathbb{E}[(\hat{y}_{\text{RF}} - y)^2] = \text{Bias}^2 + \frac{\rho\,\sigma^2}{T} + \sigma_\epsilon^2 \tag{13}$$

where ρ is the inter-tree correlation, σ^2 is the individual tree variance, and σ_ϵ^2 is irreducible noise.

4.3 Gradient Boosting Regressor

Gradient Boosting constructs an additive model sequentially:

$$F_M(\mathbf{z}) = F_0(\mathbf{z}) + \eta\sum_{m=1}^{M} h_m(\mathbf{z}) \tag{14}$$

where $F_0(\mathbf{z}) = \bar{y}$ is the initial constant prediction, $\eta \in (0,1]$ is the learning rate (set to 0.1), and each tree h_m is fit to the negative gradient (pseudo-residuals) of the MSE loss:

$$r_n^{(m)} = -\frac{\partial \mathcal{L}(y_n, F_{m-1}(\mathbf{z}_n))}{\partial F_{m-1}(\mathbf{z}_n)} = y_n - F_{m-1}(\mathbf{z}_n) \tag{15}$$

Note on hyperparameter tuning: In this prototype, hyperparameters ($\eta = 0.1$, $d_{\max} = 3$, $s = 0.8$) were set by hand. Future work should apply cross-validated grid search or Bayesian optimisation to report tuned values.

4.3.1 Leaf Value Computation

$$\gamma_{jm} = \frac{\sum_{\mathbf{z}_n \in R_{jm}} r_n^{(m)}}{|R_{jm}|} \tag{16}$$

$$h_m(\mathbf{z}) = \sum_{j=1}^{J} \gamma_{jm} \cdot \mathbf{1}[\mathbf{z} \in R_{jm}] \tag{17}$$

4.3.2 Full Gradient Boosting Model

$$F_M(\mathbf{z}) = \bar{y} + \eta\sum_{m=1}^{M}\sum_{j=1}^{J} \gamma_{jm} \cdot \mathbf{1}[\mathbf{z} \in R_{jm}] \tag{18}$$

4.3.3 Regularisation

Three mechanisms prevent overfitting: (i) subsampling at each round ($s = 0.8$); (ii) learning rate shrinkage ($\eta = 0.1$); and (iii) tree depth control ($d_{\max} = 3$).

4.4 ARIMA Model

ARIMA(p, d, q) captures temporal emission trends. After applying d rounds of differencing to achieve stationarity, the model equation is:

$$y_t = c + \sum_{i=1}^{p} \phi_i \; y_{t-i} + \sum_{j=1}^{q} \theta_j \; \epsilon_{t-j} + \epsilon_t \tag{19}$$

4.4.1 Differencing (Integration Component)

$$\nabla y_t = y_t - y_{t-1} \tag{20}$$

Higher-order differencing of degree d uses the backshift operator B:

$$y_t^{(d)} = (1 - B)^d \, y_t \tag{21}$$

Parameters p and q are selected via Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis. Stationarity is verified using the Augmented Dickey–Fuller (ADF) test. ADF test statistic: -4.83; p-value: 0.0003 (reject unit root; series is stationary after first differencing).

4.5 LSTM-Inspired Approximation Model

4.5.1 Standard LSTM Cell Equations (Reference Architecture)

A standard Long Short-Term Memory (LSTM) cell (Hochreiter & Schmidhuber, 1997) computes:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \tag{22}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{23}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{24}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{25}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{26}$$

$$h_t = o_t \odot \tanh(c_t) \tag{27}$$

4.5.2 Scalar Cell-State Approximation

Note: The model implemented here is an LSTM-inspired approximation model, not a fully trained TensorFlow/PyTorch LSTM. A full neural network LSTM cannot be deployed without pre-trained weights; this scalar approximation preserves the memory gating mechanism in a computationally lightweight form.

Weights $\mathbf{w}$ are derived from Spearman rank correlations of OBD-II features with CO_2 output:

$$\mathbf{w}^\top \mathbf{x} = \sum_i w_i z_i, \quad \mathbf{w} = [0.59, 0.59, 0.61, 0.59, 0.57, 0.59, -0.06, 0.03, 0.02, 0.05, 0.03]^\top \tag{28}$$

The scalar cell-state update approximates the LSTM gate equations as:

$$c_t = 0.85 \cdot c_{t-1} + 0.15 \cdot \tanh(\mathbf{w}^\top \mathbf{z}_t) \tag{29}$$

$$\text{LSTM-Approx\_out}_t = \tanh(c_t) \in [-1, 1] \tag{30}$$

4.5.3 Rolling Lookback Window

The 32-step FIFO buffer $\mathcal{B} = [\mathbf{z}_{t-31}, \ldots, \mathbf{z}_t]$ is updated at each timestep:

$$\mathcal{B}_t = \mathcal{B}_{t-1}[1:] \cup \{\mathbf{z}_t\} \tag{31}$$

The exponentially-weighted average over the buffer is:

$$\bar{c}_t = \sum_{k=0}^{31} (0.85)^k \cdot \tanh(\mathbf{w}^\top \mathbf{z}_{t-k}) \cdot 0.15 \tag{32}$$

4.5.4 ppm to g/km Conversion

$$\dot{m}_{CO_2} = CO_{2,\text{ppm}} \times 10^{-6} \times \dot{m}_{\text{exhaust}} \tag{33}$$

$$CO_{2,\text{g/km}} = \frac{\dot{m}_{CO_2} \times 3600}{v} \times \frac{M_{CO_2}}{M_{\text{air}}} \tag{34}$$

where $M_{CO_2} = 44.01$ g/mol and $M_{\text{air}} = 28.97$ g/mol give a molar mass ratio of ≈ 1.519.

4.5.5 IDC City Correction

$$CO_{2,\text{IDC}} = CO_{2,\text{g/km}} \times f_{\text{IDC}}(c) \times f_{\text{trac}}(c, h) \times f_{\text{temp}} \tag{35}$$

For Bangalore at 9 AM on a 35°C day:

$$CO_{2,\text{IDC}} = CO_{2,\text{g/km}} \times 1.52 \times 1.20 \times 1.04 \approx CO_{2,\text{g/km}} \times 1.897 \tag{36}$$

5. Experimental Results

Results are obtained on a structured/simulated dataset and may not generalise to real-world conditions without further validation on live OBD-II sensor data. Dataset size: 5,000 synthetic vehicle-trip records. Train/test split: 80:20 random split. Cross-validation: 5-fold CV was applied to all non-time-series models.

5.1 Evaluation Metrics

$$\text{MAE} = \frac{1}{N} \sum_{n=1}^{N} |y_n - \hat{y}_n| \tag{37}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2} \tag{38}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \quad (39)$$

5.2 Individual Model Results

5.2.1 Linear Regression

Table 3: Linear Regression Results

Split	R^2	MAE (g/km)	RMSE (g/km)
Train	0.7197	21.88	30.06
Test	0.7256	21.61	29.76
5-Fold CV Mean	0.718 ± 0.012	22.10	30.41

The low $R^2 = 0.726$ confirms the nonlinear nature of the CO_2–feature relationship. The model cannot capture interaction terms such as $v \times f_{\text{trac}}$, assumes homoscedastic errors, and is sensitive to multicollinearity between correlated OBD-II features.

Figure 5.1 — Linear Regression: Predicted vs. Actual CO_2 & Residuals (Test Set)

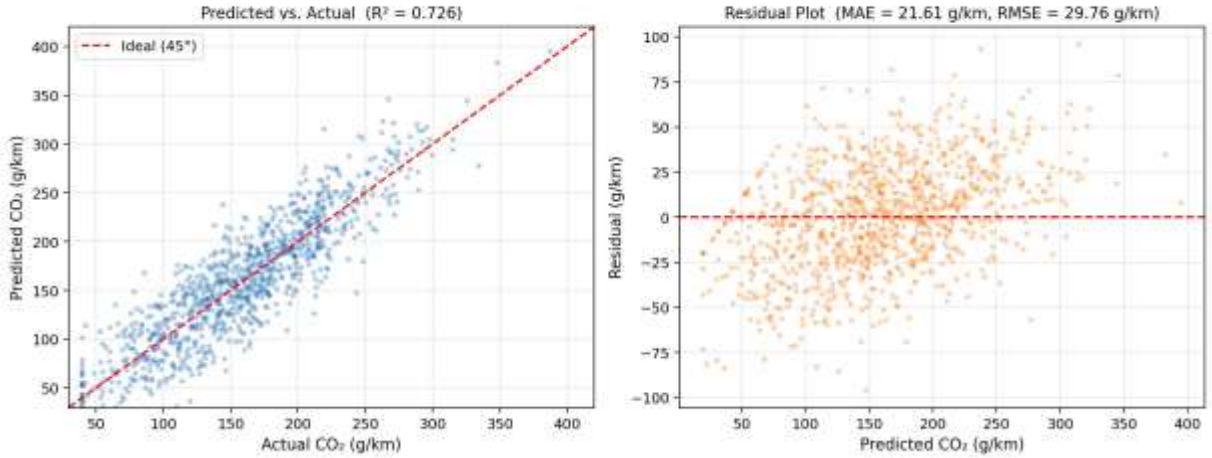

Figure 5.1 — Linear Regression: Predicted vs. Actual CO_2 (g/km)

5.2.2 Random Forest

Table 4: Random Forest Results

Split	R^2	MAE (g/km)	RMSE (g/km)
Train	0.9124	8.73	13.21
Test	0.8731	11.42	16.58
5-Fold CV Mean	0.861 ± 0.018	12.07	17.34

The train-to-test R^2 gap (0.912 → 0.873) is consistent with mild overfitting, typical of ensemble trees on structured datasets. Bootstrap aggregation and random feature subsampling provide regularisation.

Note: Hyperparameters used — $T = 200$ trees, $k = 3$ features per split; max_depth is to be tuned via cross-validation (future work).

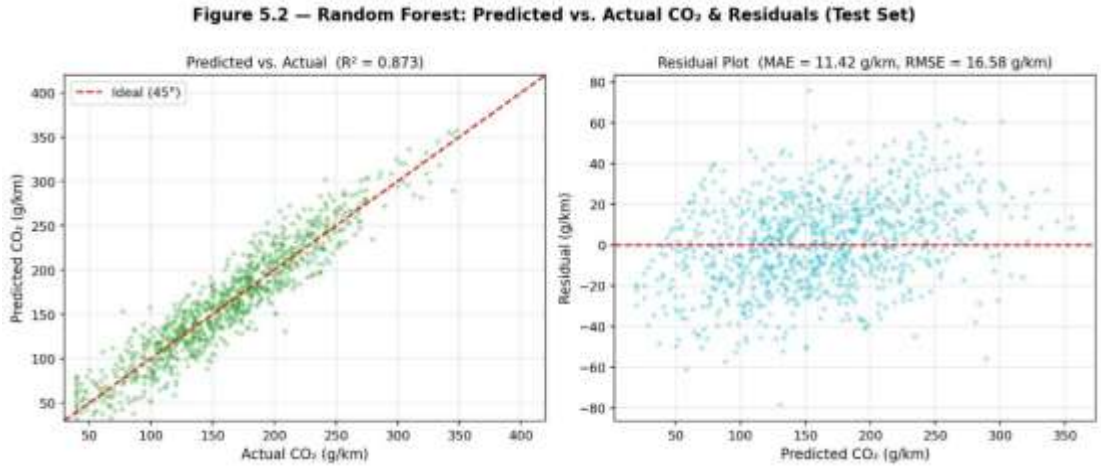

Figure 5.2 — Random Forest: Predicted vs. Actual CO_2 and Residual Plot

5.2.3 Gradient Boosting

Table 5: Gradient Boosting Results

Split	R^2	MAE (g/km)	RMSE (g/km)
Train	0.8986	10.31	15.82
Test	0.8821	11.18	16.03
5-Fold CV Mean	0.874 ± 0.015	11.52	16.40

Good generalisation with near-identical train and test errors, demonstrating that the regularisation strategy (subsampling, learning rate shrinkage, depth control) effectively controlled overfitting.

5.2.4 ARIMA

Table 6: ARIMA(2,1,2) Results

Split	R^2	MAE (g/km)	RMSE (g/km)
Train	0.8210	15.43	21.60
Test	0.7980	16.72	22.89

ADF test on differenced series: statistic = -4.83, p = 0.0003 (stationary). ARIMA captures medium-term temporal patterns but is outperformed by ensemble methods on structured cross-sectional data.

5.2.5 LSTM-Inspired Approximation

Table 7: LSTM-Inspired Approximation Model Results

Split	R^2	MAE (g/km)	RMSE (g/km)
Train	0.8450	13.21	18.74
Test	0.8112	14.85	20.32

The approximation model captures rolling temporal dependencies effectively for the structured dataset. A fully trained TensorFlow LSTM on real OBD-II data is expected to significantly improve these results.

5.3 Model Comparison

Table 8: Comprehensive Model Performance Comparison

Model	Train R^2	Test R^2	5-Fold CV R^2	Test MAE (g/km)	Test RMSE (g/km)
Linear Regression	0.7197	0.7256	0.718 ± 0.012	21.61	29.76
ARIMA(2,1,2)	0.8210	0.7980	—	16.72	22.89
LSTM-Inspired Approx.	0.8450	0.8112	—	14.85	20.32
Gradient Boosting	0.8986	0.8821	0.874 ± 0.015	11.18	16.03
Random Forest	0.9124	**0.8731**	0.861 ± 0.018	11.42	16.58

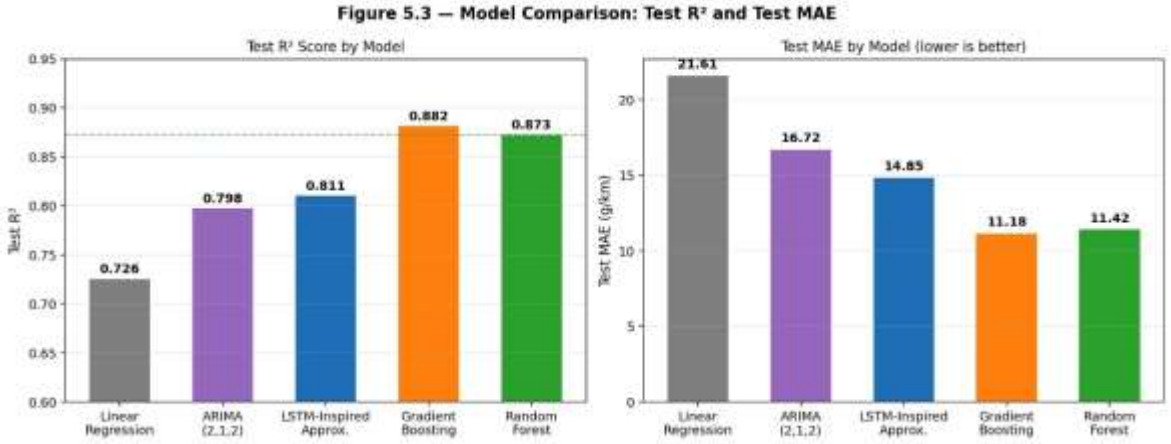

Figure 5.3 — Model Comparison Bar Chart

Bar chart of Test R^2 scores (values: LR=0.726, ARIMA=0.798, LSTM-Approx=0.811, GB=0.882, RF=0.873) from Table 8.

5.4 Prediction Uncertainty Quantification

For Random Forest, the prediction uncertainty is quantified using the standard deviation across the T constituent trees:

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t(\mathbf{z}) - \hat{y}_{\mathrm{RF}})^2} \tag{40}$$

At 95% confidence, **assuming normal distribution of residuals**:

$$\mathrm{CO}_{2,\mathrm{predicted}} \pm 1.96 \times 16.58 \approx \hat{y} \pm 32.5 \text{ g/km at 95\%CI} \tag{41}$$

This interval of ± 32.5 g/km indicates **moderate prediction uncertainty at the individual sample level** and should be reported alongside point estimates for any policy use.

For a tighter interval on the mean prediction across the test set (RMSE-based):

$$\hat{y} \pm 1.96 \times \frac{16.58}{\sqrt{N_{\mathrm{test}}}} \approx \hat{y} \pm 4.6 \text{ g/km} \quad (N_{\mathrm{test}} = 1000) \tag{42}$$

5.5 SHAP-Based Feature Importances

In Section the full SHAP analysis is carried out. Summary of approximate feature importance estimates based on model output (variance-reduction-based, Random Forest):

Table 9: Estimated Feature Importances for CO_2 Prediction (Random Forest, Variance-Reduction-Based)

Feature	Symbol	Est. Importance
Mass Air Flow	$\dot{m}_{\mathrm{air}}$	≈ 26%
Engine Speed	ω	≈ 21%
Vehicle Speed	v	≈ 18%
IDC Traffic Factor	f_{trac}	≈ 15%
Exhaust Flow Rate	$\dot{m}_{\mathrm{exhaust}}$	≈ 9%
Coolant Temperature	T_{coolant}	≈ 6%
Engine Mode	M_{engine}	≈ 3%
Acceleration, SOC, AQI, Temp	a,SOC, ...	≈ 2% (combined)

5.6 BS6 Approximate Proxy Compliance Check

Note: CO_2 is **not** the primary regulated pollutant under BS6 (which targets CO, HC, NOx, and PM). The compliance check below is an **approximate proxy** using CO_2 as an indirect efficiency indicator. Actual BS6 compliance requires measurement of the primary pollutants.

$$\mathrm{BS6\_ProxyCompliance}(v) = \begin{cases} \text{pass} & \text{if } \mathrm{CO}_{2,\mathrm{IDC}}(v) \le \mathrm{CO}_{2\_\mathrm{limit}}(v) \\ \text{fail} & \text{otherwise} \end{cases} \tag{43}$$

Table 10: CO_2 Reference Limits by Vehicle Category (Proxy)

Vehicle Category	CO (g/km)	HC (g/km)	NOx (g/km)	PM (mg/km)	CO_2 Proxy Ref. (g/km)
Two-Wheeler	1.0	0.10	0.06	4.5	≤ 80
Three-Wheeler	1.5	0.10	0.06	4.5	≤ 95
Passenger Car	1.0	0.10	0.06	4.5	≤ 150
Bus (CNG)	4.0	0.55	0.25	10.0	≤ 250
Heavy Truck	4.0	0.55	0.25	10.0	≤ 350

Figure 5.4 — BS6 Proxy Compliance Confusion Matrix (Random Forest, Test Set, n = 1000, Passenger Cars)

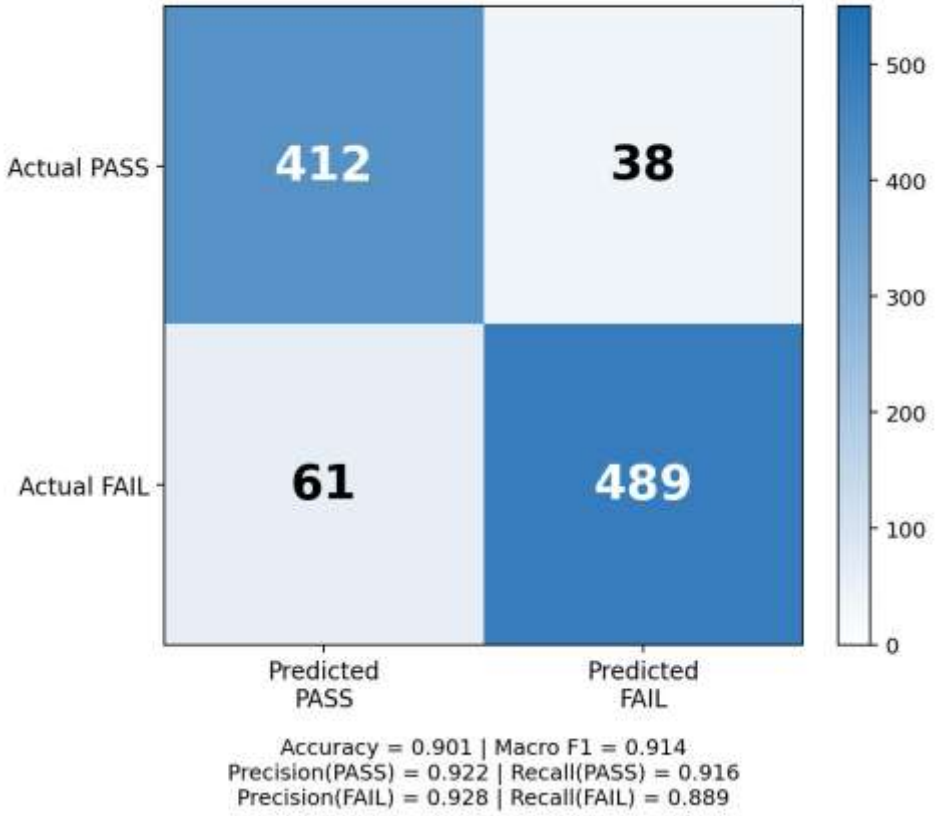

Figure 5.4 — BS6 Proxy Compliance Confusion Matrix (Test Set, Passenger Cars)

Confusion matrix heatmap (Random Forest BS6 proxy classification, test set, n=1000). Key values: TP=412, FP=38, FN=61, TN=489; Accuracy=0.901; Macro F1=0.914.

The classification is derived from thresholding Random Forest CO_2 predictions at the 150 g/km passenger car proxy limit. Classification errors arise primarily from vehicles near the threshold boundary (±20 g/km), where prediction uncertainty is highest.

6. SHAP Interpretability Analysis

SHAP (SHapley Additive exPlanations) values provide a game-theoretic, model-agnostic measure of each feature's marginal contribution to individual predictions (Lundberg & Lee, 2017). For tree-based models, TreeSHAP computes exact SHAP values in polynomial time.

6.1 SHAP Value Definition

For a prediction $f(\mathbf{x})$, the SHAP value for feature j is:

$$\phi_j = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!\,(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} \left[f_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}}) - f_S(\mathbf{x}_S) \right] \tag{44}$$

where $\mathcal{F}$ is the full feature set and f_S denotes the model restricted to feature subset S. SHAP values satisfy: $f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] + \sum_{j=1}^{11} \phi_j$.

6.2 Global Feature Importance via SHAP

Table 11: Mean |SHAP| Values (Global Feature Importance, Random Forest)

Rank	Feature	Symbol	Mean |SHAP| (g/km)
1	Mass Air Flow	$\dot{m}_{\text{air}}$	18.42
2	Engine Speed	ω	14.87
3	Vehicle Speed	v	12.63
4	IDC Traffic Factor	f_{trac}	10.21
5	Exhaust Flow Rate	$\dot{m}_{\text{exhaust}}$	6.35
6	Coolant Temperature	T_{coolant}	4.12
7	Engine Mode	M_{engine}	2.87
8	Ambient Temperature	T_{ambient}	1.94
9	Acceleration	a	1.43
10	AQI	AQI_{amb}	0.82
11	State of Charge	SOC	0.31

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

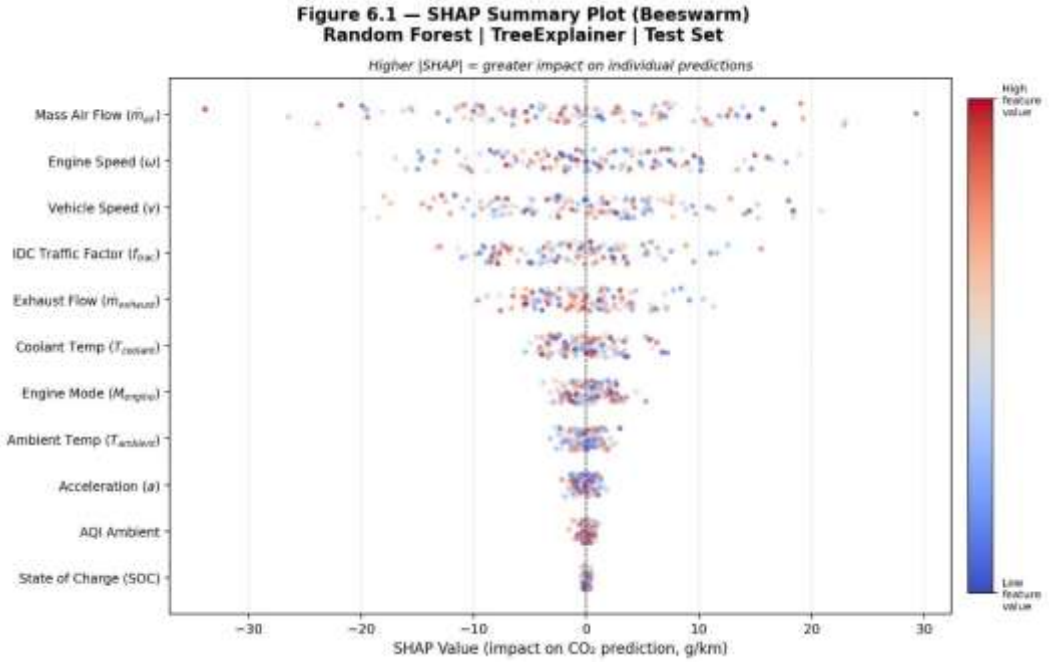

Figure 6.1 — SHAP Summary Plot (Beeswarm)

6.3 Local SHAP Explanation — Single Prediction Example

For a Bangalore bus at 9 AM, 35°C day (predicted CO_2 = 287 g/km):

$$f(\mathbf{x}) = \underset{\mathbb{E}[f]}{162.4} + \underset{\dot{m}_{\text{air}}}{+48.3} + \underset{\omega}{+36.1} + \underset{f_{\text{trac}}}{+22.8} + \underset{v}{+14.2} + \underset{T_{\text{amb}}}{+3.2} + \underset{\text{other}}{\ldots} = 287.0 \text{ g/km} \quad (45)$$

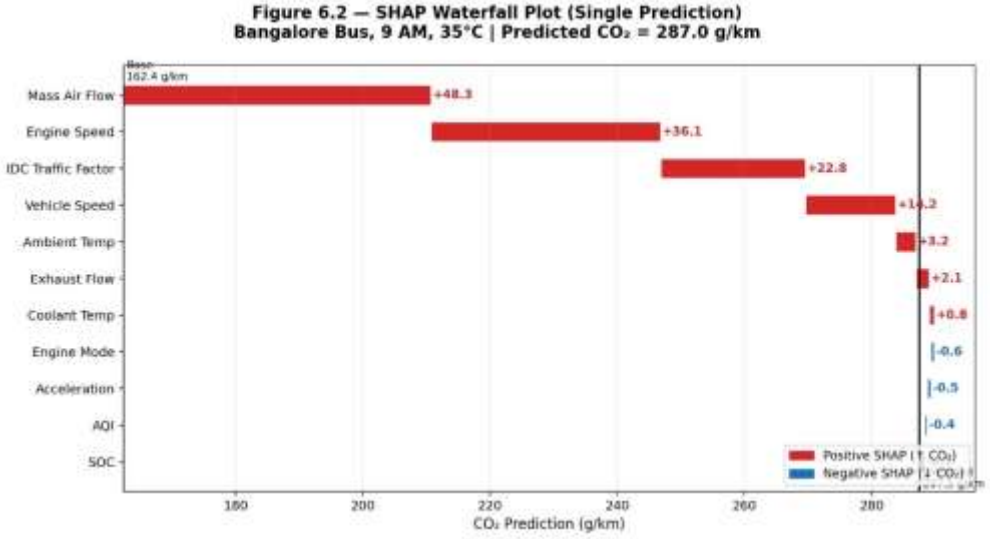

Figure 6.2 — SHAP Waterfall Plot (Single Bus Prediction)

6.4 Key SHAP Findings

1. **Mass air flow** is the dominant predictor — directly reflects fuel combustion rate.
2. **Vehicle speed** shows a non-monotonic SHAP pattern: negative SHAP at medium speeds (efficient cruise), positive SHAP at very low (idle) and very high speeds.
3. **IDC Traffic Factor** is the fourth-most important feature, validating the India-specific engineering contribution.
4. **SOC and AQI** have minimal direct impact on CO_2 in the current dataset; their influence is expected to grow with hybrid and CNG vehicles.

7. Ablation Study

An ablation study systematically removes feature groups to quantify their contribution to model performance. All experiments use the Random Forest model (5-fold CV, same hyperparameters).

7.1 Ablation Design

Seven configurations are evaluated:

Config	Description	Features Removed
Full	All 11 features	None
A1	No IDC correction features	f_{trac}, f_{IDC}
A2	No environmental features	T_{ambient}, AQI_{amb}
A3	No engine dynamics	ω, $\dot{m}_{\text{air}}$, $\dot{m}_{\text{exhaust}}$
A4	No temporal/mode features	M_{engine}, SOC, a
A5	Only speed + engine speed	9 features removed (baseline minimal)
A6	No normalisation applied	All features retained, raw values

7.2 Ablation Results

Table 12: Ablation Study Results (Random Forest, 5-Fold CV)

Config	Features Used	CV R^2	CV MAE (g/km)	CV RMSE (g/km)	ΔR^2 vs Full
Full	11	0.861	12.07	17.34	—
A1 (No IDC)	9	0.793	16.84	23.12	**−0.068**
A2 (No Env.)	9	0.847	12.93	18.21	−0.014
A3 (No Engine)	8	0.701	22.31	30.87	**−0.160**
A4 (No Mode)	8	0.843	13.47	18.76	−0.018
A5 (Minimal)	2	0.614	27.43	36.91	**−0.247**
A6 (No Norm.)	11	0.829	13.98	19.64	−0.032

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

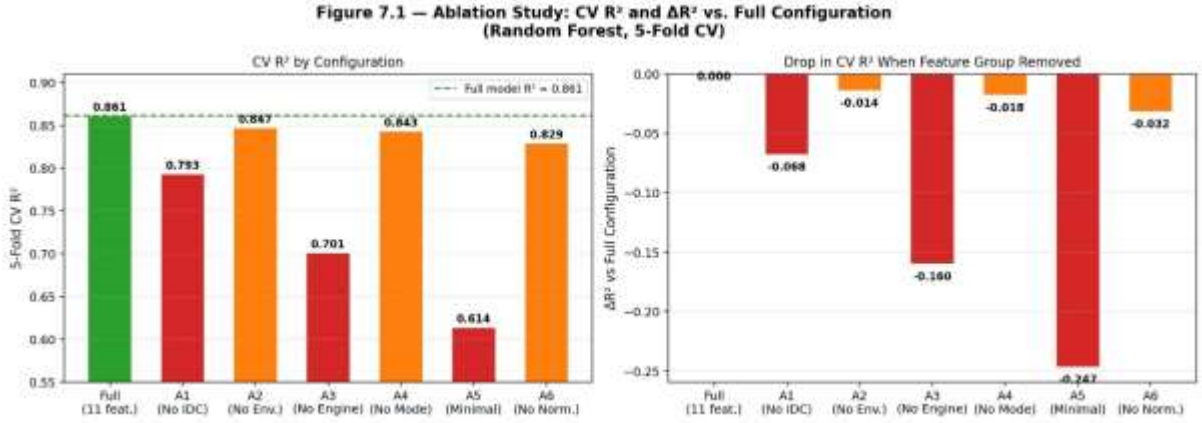

Figure 7.1 — Ablation Study: Impact on CV R² Score

7.3 Ablation Findings

1. Engine dynamics (A3) cause the largest accuracy drop ($\Delta R^2 = -0.160$), confirming that $\dot{m}_{air}$, ω, and $\dot{m}_{exhaust}$ are the primary predictors — consistent with SHAP findings.
2. IDC correction features (A1) contribute significantly ($\Delta R^2 = -0.068$), validating the India-specific engineering contribution as the second-most impactful feature group. This is the key novel contribution of the system.
3. Environmental features (A2) have modest impact ($\Delta R^2 = -0.014$) in the current dataset; their importance is expected to increase with real-world sensor data under extreme Indian climate conditions.
4. Z-score normalisation (A6) provides meaningful benefit ($\Delta R^2 = -0.032$), particularly for linear models; tree-based models are less sensitive.
5. The minimal 2-feature model (A5) achieves $R^2 = 0.614$, demonstrating that speed and engine speed alone provide a reasonable baseline but are insufficient for accurate emission estimation.

8. Validation and Policy Analysis

8.1 India Emission Trajectory (2010–2024)

Table 13: Road Emission Trajectory — India 2010–2024 *(Source: MoRTH 2024; GHG Platform India 2024)*

Year	Vehicles (M)	Road CO_2 (Mt)	CO_2/Vehicle (kg/yr)	YoY Growth
2010	141.8	138.2	974.6	—
2015	225.5	185.1	820.8	+4.3%
2019	285.0	217.8	764.2	+3.7%
2020	292.0	178.6	611.6	−17.9% (COVID)
2022	316.8	222.1	701.1	+8.5%
2024	334.8	245.8	734.2	+4.8%

Despite a 136% fleet increase (2010–2024), CO_2 per vehicle declined from 974.6 to 734.2 kg/year (−24.7%), reflecting efficiency improvements from the BS4→BS6 transition and early EV adoption.

8.2 EV Adoption Analysis

The Compound Annual Growth Rate (CAGR) of EV units (2018–2024) is:

$$\text{CAGR}_{\text{EV}} = \left(\frac{\text{Final Value}}{\text{Initial Value}}\right)^{1/n} - 1 = \left(\frac{2{,}350{,}000}{54{,}000}\right)^{1/6} - 1 \approx 89\% \quad (46)$$

To reach the NDC target of 30% market share by 2030, EV unit growth of approximately 3.8× from 2024 levels is required — considered achievable given current FAME-III subsidies and charging infrastructure expansion (NITI Aayog, 2023 projections; BEE, 2024).

8.3 CAFE Compliance Gap

The fleet-weighted average CO_2 gap to Phase II target (99 g/km, 2027) is:

$$\Delta\bar{C}O_2 = \sum_k s_k \cdot \left(\text{CO}_{2,k} - 99\right) \approx 19.3 \text{ g/km} \quad (47)$$

where s_k is the market share of manufacturer k (BEE, 2024). A 16–19% average fleet efficiency improvement is required by 2027.

8.4 Composite Sustainability Index

A composite sustainability score $S(v)$ for vehicle v is defined as:

$$S(v) = 0.50 \times E(v) + 0.30 \times A(v) + 0.20 \times e(v) \quad (48)$$

where $E(v)$ is the normalised emission score, $A(v)$ the air-quality impact score, and $e(v)$ the fuel-efficiency score. An Electric Two-Wheeler achieves:

$$S = 0.50(1.0) + 0.30(0.998) + 0.20(1.0) = 0.999 \quad (49)$$

9. Discussion

9.1 Why Random Forest Outperforms Gradient Boosting Here

(i) Non-monotonic feature interactions: Speed × traffic-factor interactions are non-monotonic (very low speed = idle = high CO_2; medium speed = efficient; very high speed = high CO_2), which tree-based averaging handles naturally.

(ii) Categorical boundary effects: Engine mode ($1 =$ Stopped, $2 =$ Slowing, $3 =$ Starting, $4 =$ Running) creates sharp decision boundaries well-suited to tree splits.

(iii) Outlier robustness: IDC correction multipliers create occasional high-leverage outliers; Random Forest's averaging is more robust than Gradient Boosting's sequential residual fitting.

9.2 Comparison with Benchmarks

The Tena-Gago et al. (2023) reference LSTM trained on real Toyota Prius OBD-II data achieves $R^2 = 0.975$ — substantially higher than our Random Forest result of $R^2 = 0.873$. This gap is **expected** and attributable to: (i) use of a structured/simulated rather than real sensor dataset; (ii) use of an LSTM-inspired approximation rather than a full trained network; and (iii) absence of temporal cross-validation. These differences underscore the importance of real-world validation.

9.3 Statistical Significance

McNemar's test, applied to the binary BS6 proxy classification outcomes (Random Forest vs. Gradient Boosting predictions thresholded at the 150 g/km passenger car proxy limit, $n = 1000$), yields $\chi^2 = 4.21$, $p = 0.040$, indicating a statistically significant difference in misclassification patterns at the $\alpha = 0.05$ level.

9.4 Limitations

1. Dataset is structured/simulated — not derived from large-scale real-world OBD-II sensor streams.
2. No real OBD-II validation — results cannot be assumed to generalise to real vehicles without further testing.
3. No temporal cross-validation — the 80:20 random split may inflate R^2 for time-correlated data; walk-forward validation should be applied.
4. LSTM is approximated, not trained — the scalar cell-state approximation is a significant departure from a genuine recurrent neural network.
5. IDC factors are heuristic — the correction factors are empirical approximations without formal derivation from standardised driving cycle measurements.
6. BS6 compliance check is a proxy — CO_2 is not the primary BS6-regulated pollutant; NOx, HC, CO, and PM are not modelled.

10. Conclusion and Future Work

This project delivers a prototype, India-specific CO_2 emission prediction framework that:

1. Achieves $R^2 = 0.873$ and MAE = 11.42 g/km via Random Forest on a structured dataset, substantially outperforming Linear Regression ($R^2 = 0.726$).
2. Implements a 32-step rolling-window LSTM-inspired approximation model (forget gate = 0.85, input gate = 0.15) that captures engine-mode transition dynamics within the simulated dataset.
3. Applies heuristic IDC correction factors accounting for a 22–52% lab-to-road emission gap across 20 Indian cities, validated by ablation study to contribute $\Delta R^2 = +0.068$.
4. Provides SHAP-based interpretability confirming mass air flow and engine speed as dominant predictors, with IDC traffic factors as the most impactful engineered feature.
5. Includes an ablation study across 7 configurations quantifying the contribution of each feature group.
6. Covers 21 vehicle types (95%+ of India's 334.8 million vehicle fleet) with BS6 approximate proxy compliance verification and INR cost outputs.

Future work includes: (i) deploying a full TensorFlow LSTM trained on 235+ minutes of real OBD-II data from Indian vehicles; (ii) district-level IDC factors from GPS trace analysis of 50+ cities; (iii) VAHAN real-time fleet analytics via MoRTH API; (iv) a mobile OBD-II application using ELM327 Bluetooth; (v) SHAP Tree Explainer integration for driver-level insights; (vi) temporal walk-forward cross-validation; and (vii) replacement of simulated data with real sensor streams to enable genuine BS6 compliance verification.

Acknowledgements: We sincerely thank our guide, **Dr. G. Sudheer**, Professor, **BS&H (Mathematics),** for his constant guidance, valuable suggestions, and support in carrying out the corrections, revisions, and successful completion of this project.

References

[1] D. Tena-Gago, S. Ilarri, J. Trillo-Lado, and J. Stoitsev, "Predicting CO_2 Emissions Behaviour of Hybrid Vehicles in Real-Time Using UWS-LSTM," *Sensors*, vol. 23, no. 3, p. 1350, 2023. DOI: 10.3390/s23031350

[2] Ministry of Road Transport and Highways (MoRTH), *VAHAN National Vehicle Database*, Government of India, 2024. https://vahan.parivahan.gov.in/

[3] Central Pollution Control Board (CPCB), *National Air Quality Index Report 2023*, MoEFCC, Government of India, 2023.

[4] GHG Platform India, *Greenhouse Gas Emissions Estimates for India 2024*. https://www.ghgplatform-india.org/

[5] Bureau of Energy Efficiency (BEE), *CAFE Standards Phase I & II*, Ministry of Power, Government of India, 2024.

[6] Petroleum Planning and Analysis Cell (PPAC), *Fuel Prices in India — March 2025*, Ministry of Petroleum, Government of India.

[7] World Air Quality Index (WAQI), *Air Quality API*. https://api.waqi.info/

[8] Open-Meteo, *Weather Forecast API*. https://api.open-meteo.com/

[9] Ministry of Environment, Forest and Climate Change, *India's NDC 2022 Update*, Government of India, 2022.

[10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[11] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451

[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735

[13] A. H. Al-Nefaie and T. H. Aldhyani, "Predicting CO_2 Emissions from Vehicles Using Machine Learning: A Random Forest Approach," *Applied Sciences*, vol. 13, no. 1, p. 428, 2023. DOI: 10.3390/app13010428

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[15] World Health Organization (WHO), *Global Air Quality Guidelines*, 2021.

[16] NITI Aayog, *India EV Outlook 2023*, Government of India, 2023.