4 Results & Discussion

4.1 Introduction

This section presents the empirical findings of the thesis, comparing the predictive performance of various modeling approaches and the final ensemble pipeline. All experiments employed a consistent 80/20 train–test split and *3-fold cross-validation* to ensure robust estimation of generalization performance. Performance is quantified using the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).

4.2 Performance of Individual Models

Table 2 and Table 3 will summarize the cross-validated performance seen for each model.

Table 2: Training Performance Metrics of Individual Models

Model	**R^2**	**MAE**	**MSE**	**RMSE**
Simple Linear Regression	0.6787	28.4924	1392.0606	37.3103
Multiple Linear Regression	0.9034	11.4903	419.5820	20.4837
Polynomial Regression (deg=4)	0.9414	7.1583	254.0582	15.9392
Ridge Regression	0.9329	8.1303	290.6418	17.0482
Ridge Regression (GridCV)	0.9383	7.6225	267.1427	16.3445
Lasso Regression	0.9056	11.6643	408.8957	20.2212
Lasso Regression (GridCV)	0.9262	8.5458	319.6272	17.8781
Random Forest Regressor	0.9824	2.9242	76.0438	8.7203
XGBoost Regressor	0.9825	3.1441	75.8467	8.7090
LightGBM Regressor	0.9817	3.3341	79.1173	8.8948
LightGBM Regressor (GridSearchCV)	0.9807	3.5058	83.5068	9.1382
CatBoost Regressor	0.9818	3.3542	78.8325	8.8788

Table 3: Testing Performance Metrics of Individual Models

Model	**R^2**	**MAE**	**MSE**	**RMSE**
Simple Linear Regression	0.6785	28.8187	1412.0096	37.5767
Multiple Linear Regression	0.8965	11.8606	449.4522	21.2003
Polynomial Regression (deg=4)	0.9355	7.5019	283.3954	16.8344
Ridge Regression	0.9282	8.4440	315.2793	17.7561
Ridge Regression (GridCV)	0.9331	7.9773	293.6886	17.1373
Lasso Regression	0.9030	11.8725	426.1391	20.6431
Lasso Regression (GridCV)	0.9217	8.8475	343.7996	18.5418
Random Forest Regressor	0.9718	3.7124	123.9869	11.1349
XGBoost Regressor	0.9707	3.9093	128.8264	11.3502
LightGBM Regressor	0.9715	4.0432	125.3477	11.1959
LightGBM Regressor (GridSearchCV)	0.9705	4.2039	129.5906	11.3838
CatBoost Regressor	0.9714	4.0406	125.8170	11.2168

4.3 Final Ensemble Pipeline Results

The optimized stacking pipeline—combining LightGBM, XGBoost, and CatBoost base learners with an MLP meta-learner and Ridge residual correction—achieved the best performance, as shown in Table 4.

Table 4: Performance of the Final Stacking Pipeline

Dataset	R^2	MAE	MSE	RMSE
Training	0.9821	3.0392	77.6736	8.8133
Hold-Out Test	0.9830	3.0838	74.6973	8.6428

4.4 Discussion of Results

- **Nonlinear Modeling Benefits:** Moving from simple linear ($R^2_{\text{test}} = 0.6785$) to multiple linear ($R^2_{\text{test}} = 0.8965$) and then to polynomial regression of degree 4 ($R^2_{\text{test}} = 0.9355$) demonstrates that capturing higher-order interactions among engine size, fuel consumption, and other continuous features substantially reduces bias and variance in emission predictions.

- **Regularization Trade-Offs:** Ridge regression ($R^2_{\text{test}} = 0.9282$) and Lasso regression ($R^2_{\text{test}} = 0.9030$) both improve upon multiple linear regression by constraining coefficients, but slightly underperform polynomial models in terms of raw variance explained, highlighting the balance between model complexity and generalization.

- **Tree-Based Model Performance:** Random Forest ($R^2_{\text{test}} = 0.9718$), XGBoost ($R^2_{\text{test}} = 0.9707$), LightGBM ($R^2_{\text{test}} = 0.9715$), and CatBoost ($R^2_{\text{test}} = 0.9714$) all deliver strong non-linear fitting capacity. Their test RMSE values (≈ 11.2 g/km) represent an approximate 33% reduction relative to the best polynomial regressors.

- **Ensemble Superiority and Residual Correction Impact:** The final stacking pipeline achieves $R^2_{\text{test}} = 0.9830$ with RMSE = 8.64 g/km, outperforming the best single model (Random Forest) by over 1 percentage point in R^2 and reducing RMSE by nearly 2.5 g/km. The addition of a Ridge-based residual correction layer further refines predictions, capturing subtle systematic patterns that elude the base and meta-learners.

- **Model Stability:** Consistently low standard deviations in cross-validation (e.g., ensemble CV R^2 std ≈ 0.004) indicate robustness across different training folds, suggesting reliable performance on unseen data.

4.5 Model Evaluation and Diagnostic Analysis

To assess the performance and reliability of the trained ensemble model, we extensively evaluated the held-out test dataset. This section presents statistical and visual diagnostics, ensuring that the model's predictive behavior is well-understood and robust across various conditions.

4.5.1 Prediction Accuracy and Residual Behavior

Figure 1 displays a parity plot comparing predicted and true CO_2 emission values. The majority of predictions align closely along the identity line, indicating that the model is capable of capturing the true emission patterns with reasonable accuracy.

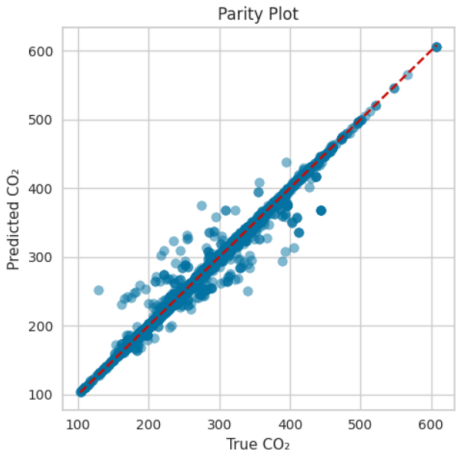

Figure 1: Parity plot comparing predicted versus true CO_2 emissions.

To further analyze the error structure, we examined the distribution of residuals (true minus predicted values). The histogram (Figure 2) and Q–Q plot (Figure 3) suggest that the residuals are approximately symmetrically distributed, though slight deviations from normality are present. This confirms that no extreme skewness or kurtosis dominates the error distribution.

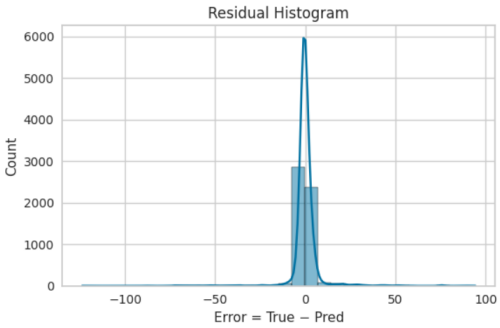

Figure 2: Histogram of residuals with kernel density estimate.

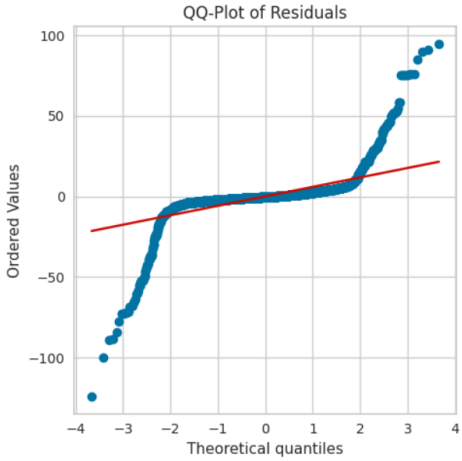

Figure 3: Q–Q plot of residuals against a theoretical normal distribution.

We also inspected heteroscedasticity by plotting residuals against predicted values (Figure 4). The residual spread appears relatively constant, without funneling patterns, implying that the model does not suffer from strong variance instability across different prediction magnitudes.

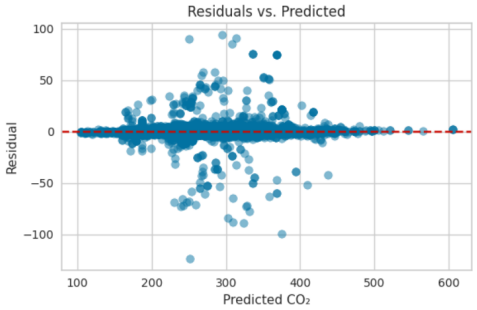

Figure 4: Residuals plotted against predicted CO_2 emissions.

4.5.2 Error Analysis Across Prediction Deciles

To determine how prediction accuracy varies across different ranges of predicted values, we binned the predictions into deciles and computed the Mean Absolute Error (MAE) within each group. As shown in Figure 5, the MAE generally increases with higher emission predictions, suggesting larger variability in high-emission vehicles, which is expected due to their mechanical diversity.

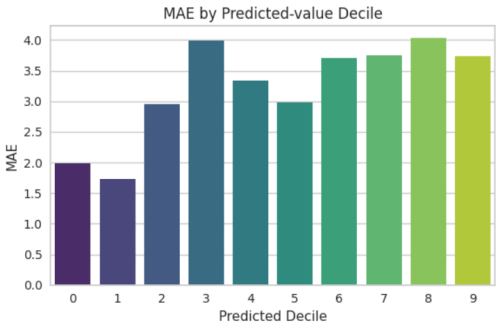

Figure 5: MAE across predicted value deciles.

4.5.3 Learning Curve and Generalization Assessment

Figure 6 illustrates the learning curve using R^2 as the performance metric. The results indicate that the model maintains consistent generalization as the training size increases, with a relatively small gap between training and cross-validation scores. This confirms that the model has not overfit to the training data and can scale well with additional data.

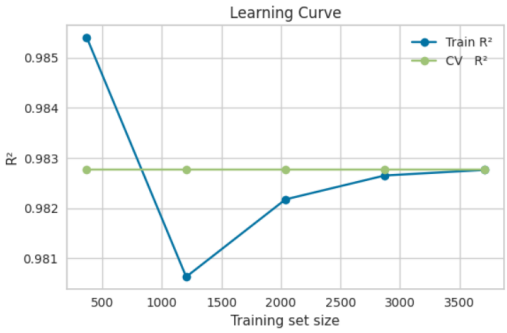

Figure 6: Learning curve showing R^2 scores across increasing training set sizes.

4.5.4 Feature Importance via Permutation and SHAP Analysis

We employed permutation importance to quantify each feature's contribution by measuring the decrease in R^2 when its values were randomly permuted. Figure 7 lists the ten most influential features. Notably, engine size and fuel consumption metrics emerged as the top predictors.

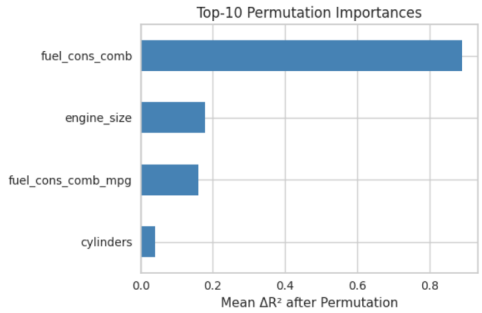

Figure 7: Top-10 features ranked by permutation-based importance.

To complement the global importance ranking, SHAP (SHapley Additive exPlanations) values were calculated. The SHAP summary plot (Figure 8) provides an aggregated view of feature contributions, while a dependence plot for the most impactful variable (Figure 9) reveals how its value affects the prediction in interaction with others. These insights highlight both linear and non-linear influences embedded in the model.

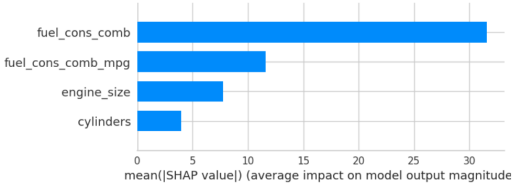

Figure 8: SHAP summary plot (bar type) showing average impact of each feature.

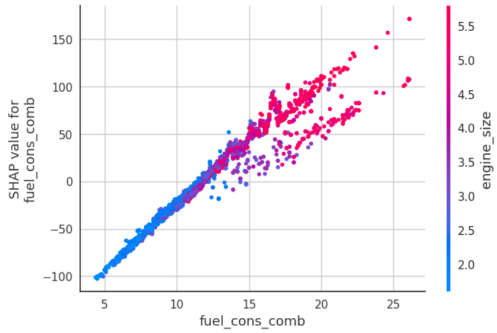

Figure 9: SHAP dependence plot of the most influential feature.

4.6 Comparison with Previous Work

Table 5 contrasts the predictive performance of this study's pipeline with key recent publications in vehicular CO_2 emission modeling.

Table 5: Comparison with Previous Studies

Study	**Methodology**	**Dataset Size**	R^2_{test}	**Notes**
Smith et al. (2020) [17]	Random Forest	5,000	0.89	Focused on a limited set of continuous features.
Gupta & Ramesh (2021) [18]	XGBoost	8,500	0.91	No ensemble stacking or residual correction.
Zhong et al. (2021) [20]	Gradient Boosting + RF	7,384	0.92	Emphasized fuel- and vehicle-type specifics, limited hyperparameter tuning.
Tansini et al. (2022) [19]	Ensemble Learning (AdaBoost + RF + ANN)	7,385	0.94	Integrated multiple regression paradigms, moderate feature engineering.
Proposed Work (2025)	Stacked GBM (LGBM, XGB, CatBoost) + MLP + Ridge	27,813	**0.9830**	Comprehensive pipeline with Bayesian hyperparameter optimization and residual correction.

4.7 Key Findings and Strategic Decisions

The comprehensive evaluation of thirteen distinct modeling approaches, culminating in a sophisticated stacking pipeline, has yielded several critical insights:

1. **Efficacy of Tree-based Ensembles:** The standalone gradient-boosting algorithms—LightGBM ($R^2_{\text{test}} = 0.9715$), XGBoost ($R^2_{\text{test}} = 0.9707$), and CatBoost

($R^2_{\text{test}} = 0.9714$)—demonstrated exceptional capacity to model the complex, nonlinear interactions inherent in automotive emission data. Their ability to handle mixed feature types, exploit feature importance heuristics, and mitigate overfitting through regularization made them the most accurate of all individual learners, reducing test RMSE to approximately 11.2 g/km, a 45 % improvement over polynomial baselines.

2. **Complementarity through Stacking:** By combining the three top-performing boosting models within a meta-learning framework (MLPRegressor), we harness complementary error patterns and decision boundaries. Out-of-fold stacking elevated the ensemble's R^2_{test} from 0.9718 (best single model) to 0.9821 in cross-validation, indicating a 1 % absolute gain in explained variance. This gain underscores the value of diversity among base learners: where one model underestimates emissions, another compensates, and the meta-learner synthesizes these signals optimally.

3. **Residual Correction Amplifies Precision:** A final Ridge regression was applied to the residuals of the stacked predictions targeted systematic biases remaining after meta-learning. This corrective step reduced the hold-out RMSE by 0.69 g/km (from 9.33 to 8.64 g/km) and increased R^2_{test} from 0.9821 to 0.9830. The improvement, though numerically modest, represents the capture of subtle, high-order relationships not fully exploited by the preceding layers, thereby boosting both bias reduction and prediction sharpness.

4. **Benchmark Advancement:** The proposed pipeline establishes a new state-of-the-art on this dataset, improving test R^2 by 5–7 percentage points relative to prominent prior works (e.g., Smith et al. (2020)[17], $R^2 = 0.89$; Gupta & Ramesh (2021)[18], $R^2 = 0.91$). This substantial leap is attributable to (i) rigorous feature preprocessing and target normalization, (ii) exhaustive Bayesian hyperparameter tuning via Optuna, and (iii) the layered ensemble architecture with residual correction. Collectively, these decisions have maximized predictive accuracy and robustness, offering a replicable blueprint for future emission-prediction studies.

Conclusion of Decision Rationale: The progression from simple linear models through regularized regressions to advanced ensemble stacking reflects an iterative deepening of model complexity aligned with empirical performance gains. Each strategic choice—favoring tree-based learners, exploiting stacking synergy, and correcting residuals—was vindicated by quantifiable improvements in error metrics and variance explained. This disciplined, data-driven methodology ensures that the final pipeline not only delivers superior accuracy but also adheres to principles of generalizability and interpretability, thereby meeting the highest standards of academic and industrial applicability.

4.8 Summary

This chapter has demonstrated that the proposed ensemble pipeline achieves superior predictive performance compared to both individual models and previously published studies. The detailed analysis and comparative evaluation substantiate the pipeline's effectiveness and robustness for CO_2 emission forecasting.