FMDB ***Transactions on Sustainable Environmental Sciences***

VayuCredit: A Data-Driven and Machine Learning Framework for Carbon Emission Monitoring and Credit Quantification Using Ensemble Learning

Dharni Patel[1,*], Urvi Deore[2], Y. Anantha Vishwa Priya[3], M. Mohamed Sameer Ali[4], Luigi Orlando Freire Martínez[5], Jaime Alfonso Flores Navas[6]

[1,2,3]Department of Artificial Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[4]Department of Computer Science and Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[5]School of Engineering and Applied Sciences, Technical University of Cotopaxi, Latacunga, Cotopaxi, Ecuador.
[6]Faculty of Sciences, National Autonomous University of Mexico, Mexico City, Ciudad de México, Mexico.
dharnipatel2006@gmail.com[1], urvideore151@gmail.com[2], ananthavishwapriya@gmail.com[3], sameerali7650@gmail.com[4], luigi.freire@utc.edu.ec[5], jaguarnegro175@gmail.com[6]

Abstract: The economic development sector's industrial operations have struggled with carbon emission monitoring and regulatory compliance, especially as India prepares to implement its Carbon Credit Trading Scheme (CCTS) by October 2026. VayuCredit analyses Indian companies' carbon emissions in different CCTS-regulated regions using three machine learning models. The World in Data Global Carbon Project trained all three models: Scale-invariant classifiers predict emission trends and make CCTS-aligned recommendations, Copula-Based Outlier Detection models find unsupervised anomalies, and XGBoost regressors with log-transformed targets and Optuna hyperparameters. From 1990 to 2022, this collection has 7,748 records from 235 nations. Researchers avoid cross-country data leakage by generalising across unobserved national emission trends using GroupKFold cross-validation. VayuCredit's real-time CCTS compliance module meets MoEFCC's October 2025 and January 2026 GEI targets for cement, aluminium, chlor-alkali, pulp and paper, petroleum refining, petrochemicals, and textiles. GHG Protocol Tier 1 emission parameters for India are based on CEA 2023 grid data and BEE benchmark. The compliance engine calculates Indian Rupee profits and penalties for Carbon Credit Certificates (CCCs) at a market price of ₹830-1,000 per tonne CO_2e. Double market-rate BEE fines. Emission computation, ML prediction, anomaly detection, trend recommendation, and CCTS compliance evaluation are FastAPI REST API endpoints. All three multi-models combined national-scale training data with company-scale industrial inference to exceed accuracy standards in industrial carbon monitoring and compliance management testing.

Keywords: Anomaly Detection; Unsupervised Learning; Random Forest; Carbon Emission Prediction; Carbon Credit Trading Scheme (CCTS); Economic Development; Carbon Credit Certificates (CCCs).

Received on: 09/03/2025, **Revised on:** 16/05/2025, **Accepted on:** 05/08/2025, **Published on:** 03/03/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSESS

DOI: https://doi.org/10.69888/FTSESS.2026.000693

Cite as: D. Patel, U. Deore, Y. A. V. Priya, M. M. S. Ali, L. O. F. Martínez, and J. A. F. Navas, "VayuCredit: A Data-Driven and Machine Learning Framework for Carbon Emission Monitoring and Credit Quantification Using Ensemble Learning," *FMDB Transactions on Sustainable Environmental Sciences*, vol. 3, no. 1, pp. 43–53, 2026.

[*]Corresponding author.

1. Introduction

Climate change driven by anthropogenic greenhouse gas emissions represents one of the most pressing global challenges of the twenty-first century [21]. The Intergovernmental Panel on Climate Change requires that limiting global warming to 1.5 degrees Celsius above pre-industrial levels requires a 45% reduction in global carbon dioxide emissions by 2030, relative to 2010 levels. Industrial operations, including energy generation, manufacturing, transportation, and cement production, collectively contribute around 73% of global GHG emissions [17]. As nations worldwide implement carbon pricing mechanisms and emissions trading systems, industries face increasing pressure to monitor, report, and reduce their carbon footprint with precision and accountability. India, as the third-largest emitter of carbon dioxide globally, has made significant commitments to reduce carbon emissions [16]. The nation has pledged to achieve 500 gigawatts of non-fossil fuel energy capacity by 2030 and to reach almost zero net emissions by 2070. An important policy instrument in this trajectory is the Carbon Credit Trading Scheme, notified in June 2023 and made operative in April 2025 under the purview of the Bureau of Energy Efficiency [2].

The CCTS currently covers six industrial sectors, including cement, aluminium, chlor-alkali, pulp and paper, iron and steel, and petroleum refining, with additional sectors phased in over subsequent compliance cycles. Regulated entities are assigned legally binding Greenhouse Gas Emission Intensity targets, and carbon credit trading is expected to be fully operational by October 2026. Non-compliant entities face penalties equivalent to twice the prevailing average market credit price [3]. Despite the urgency of industrial carbon monitoring, small and medium enterprises in India often lack the technical infrastructure and expertise needed for accurate emission calculations, trend analysis, and regulatory compliance assessments. Existing carbon accounting tools are designed only for large corporations in developed economies and do not address the specific requirements of small-scale Indian industries operating under the CCTS framework [4]. Furthermore, the heterogeneous nature of global emission data, where annual CO2 output ranges from under 0.001 million tonnes for small island nations to over 12,000 million tonnes for major emitters, creates significant challenges for machine learning models attempting to learn universal emission patterns. Machine learning approaches for carbon emission analysis have attracted considerable research interest in recent years [5].

Ensemble methods such as Random Forest, Gradient Boosting, and XGBoost have demonstrated effectiveness in predicting national and regional emission levels [18]. However, most existing studies focus exclusively on prediction tasks without addressing the complementary challenges of anomaly detection and regulatory compliance integration. Anomaly detection in emission data has received comparatively little attention. Traditional Isolation Forest approaches exhibit fundamental limitations when applied to emission datasets characterised by highly correlated features such as coal, oil, gas, and energy consumption [23]. Statistically grounded alternatives such as ECOD and COPOD from the PyOD library offer more principled distributional detection without relying on random feature partitioning [20]. This paper presents VayuCredit, a comprehensive AI-powered carbon emission monitoring platform that addresses three interconnected challenges through a unique system: emission prediction via an optimised gradient boosting regressor, unsupervised anomaly detection using copula-based distributional methods, and emission trend classification for proactive compliance decision-making [6]. The platform also integrates India-specific CCTS compliance features that are directly grounded in official BEE sector targets and penalty rules [22]. The primary contributions of this work are as follows:

- An XGBoost emission prediction model trained on 7,748 records across 235 countries with 25 engineered features, Optuna hyperparameter tuning, and GroupKFold cross-validation, achieving $R^2 = 0.9923$ and SMAPE = 4.40 percent.
- A COPOD-based anomaly detector with per-country z-score normalisation that achieves an AUC of 0.7152, outperforming Isolation Forest by 2.2 percent with markedly improved precision.
- An XGBoost trend classifier using scale-invariant ratio and percentage features, achieving 89.81% accuracy and an AUC of 0.9562, validated against both country-scale and Indian SME-scale inputs.
- A FastAPI-served backend with four India-specific CCTS compliance modules grounded in real BEE sector targets, GEI calculations, and penalty rules from the June 2023 regulatory notification.

2. Literature Review

2.1. Carbon Emission Prediction Using Machine Learning

The application of machine learning to carbon emission forecasting has evolved substantially over the past period. Early approaches relied on statistical methods such as ARIMA and exponential smoothing models, which assume linear time series and are effective for short-term, single-entity forecasting. These methods struggle to capture nonlinear relationships between economic activity, energy consumption patterns, and resulting emissions. Ensemble tree-based methods marked a significant advancement. Mardani et al. [19] identified that Random Forest and Gradient Boosting are consistently strong performers across multiple national emission datasets. Liang and Liu [8] applied XGBoost regression to predict provincial-level emissions in China, achieving an R-squared value exceeding 0.95. However, the authors noted that random train-test splitting in country-level data can also lead to information leakage when temporal patterns are present. A critical methodological improvement is the use of GroupKFold cross-validation, which keeps all observations from a given country within the same fold, thereby preventing cross-entity leakage. The Optuna hyperparameter optimisation framework further helps enable efficient search over large configuration spaces without exhaustive grid search. Deep learning architectures, including LSTM and Bidirectional LSTM networks, have also been applied to emission time series. Li et al. [12] demonstrated that LSTM models outperform traditional regression approaches by 15-20% in root-mean-square error on multi-step forecasting tasks. Wang et al. showed that BiLSTM architectures achieve particularly strong performance when the structural breaks create asymmetric temporal dependencies [7]. However, the computational overhead of sequence-based architectures and their sensitivity to sequence length present practical challenges for country-level datasets with relatively short annual time series, making well-tuned gradient boosting a competitive and often preferable alternative.

2.2. Anomaly Detection in Environmental Data

Anomaly detection in emission data serves many purposes, such as identifying data quality issues, flagging equipment malfunctions, detecting unusual operational patterns, and highlighting geopolitical events that disrupt emissions. The Isolation Forest algorithm has been used for unsupervised anomaly detection because of its computational efficiency [9]. However, it operates by random partitioning of feature space, which eventually makes it inherently weak when features are highly correlated, as it is typical in emission datasets where coal, oil, gas, and total energy consumption move together [26]. The PyOD library provides a comprehensive collection of outlier detection algorithms. Among these, the ECOD model models the distribution of each feature independently using empirical cumulative distribution functions, identifying observations that fall in the extreme tails across multiple features [10]. COPOD extends this approach by modelling inter-feature dependencies through copula functions. A fundamental challenge in evaluating unsupervised detectors is the absence of ground truth labels; synthetic anomaly injections provide a principled evaluation framework, though the injection protocol must generate anomalies representative of actual operational disruptions to yield meaningful performance estimates [24].

2.3. Emission Classification and Recommendation Systems

Classification of emission patterns enables automated recommendation generation. Conventional approaches classify entities as high or low emitters based on absolute emission levels, which can yield artificially inflated accuracy when features directly encode the target label [11]. A more practically valuable formulation uses relative measures: determining whether a given entity is above its historical media and on an upward trajectory, rather than whether its absolute volume is large. Random Forest classifiers have demonstrated strong performance on such tasks due to their resistance to overfitting and ability to handle mixed feature types [13].

2.4. Carbon Credit Trading and Compliance Systems

Carbon credit trading systems have been implemented in over 45 jurisdictions globally, with the EU Emissions Trading System serving as the most established example [25]. India's CCTS represents the nation's first compliance carbon market. The GEI metric, defined as total CO_2 emissions divided by annual production output, serves as the compliance benchmark, ensuring that emission intensity rather than absolute volume determines compliance status [14]. Digital platforms for CCTS compliance management remain nascent in the Indian context. Existing commercial solutions do not incorporate CCTS-specific calculations, sector targets, or the penalty mechanisms prescribed by the Indian regulatory framework [15]. VayuCredit seeks to address this gap.

3. Proposed Methodology

This section presents the complete VayuCredit framework, encompassing the dataset, preprocessing pipeline, feature engineering, the three machine learning models, and the CCTS compliance features.

.

3.1. Dataset Description

The training dataset is sourced from the Our World in Data Global Carbon Project (2024 edition), which aggregates data from the Global Carbon Project, Climate Watch, and the International Energy Agency (Table 1).

Table 1: Dataset characteristics

Parameter	Value
Source	Our World in Data — Global Carbon Project
Time Period	1990 to 2022

Countries	235 (aggregate regions excluded)
Records (after filtering)	7,748
Raw Features	79 columns (20 selected for analysis)
Engineered Features	25 total (5 raw + 20 derived)
Target (Model 1)	total_co2_mt in Megatonnes CO_2 (log-transformed during training)
Target Range	0.000 to 245.7 Mt
Target (Model 3)	Emission trend class: high_emission / low_emission (binary)
License	Creative Commons BY 4.0

The dataset is licensed under Creative Commons BY 4.0 and freely available for research purposes. The raw dataset contains 50,411 rows and 79 columns. After filtering to retain only the 1990-2022 period and removing non-sovereign aggregate entities such as the World, Asia, and the EU-27, the working dataset comprises 7,748 records across 235 countries.

3.2. Data Preprocessing and Feature Engineering

Missing values are addressed through a multi-strategy imputation pipeline. Forward-fill and backwards-fill are applied within each country group to leverage temporal continuity, followed by global median imputation for residual missing values. Columns with more than 60 percent missing values dropped entirely. Duplicate country-year records are removed, retaining the first occurrence. Outlier capping using Winsorisation at three times the interquartile range is applied to key emission columns rather than removing records; in total, 1,067 outliers are capped in total CO2, 1,239 in coal emissions, 784 in oil emissions, and 1,208 in gas emissions. Feature engineering produces 25 features in total, combining 5 raw inputs with 20 derived variables. Temporal features include year-over-year percentage change, absolute year-over-year change, one-, two- and three-year lag values, and three- and five-year rolling averages. Ratio features include fuel share percentages for coal, oil, and gas; the ratio of current emissions to the previous year and to the three-year average; and the rolling mean relative to the country median. Economic features include emission intensity per unit of GDP. Log-transformed versions of the four main raw columns are also included to address scale heterogeneity across the dataset.

3.3. Model 1: XGBoost Emission Predictor

The emission prediction task is formulated as a supervised regression problem. The target variable is total CO2 in megatonnes, ranging from 0.000 to 245.7 Mt across the filtered dataset. A log transformation is applied to the target during training to address the extreme range, and predictions are inverse-transformed for evaluation. A RobustScaler, which uses the median and interquartile range rather than the mean and standard deviation, is applied to input features to prevent distortion from extreme values. Figure 1 Model 1-Emission Predictor performance. (Left) Actual vs Predicted CO_2 values on test set (R^2=0.9976). (Centre) Residual plot showing prediction errors are small and randomly distributed around zero, indicating no systematic bias. (Right) Top 12 feature importances- temporal lag features (co2_lag_1yr, co2_rolling_3yr) dominate, confirming that the historical emission trajectory is the strongest predictor.

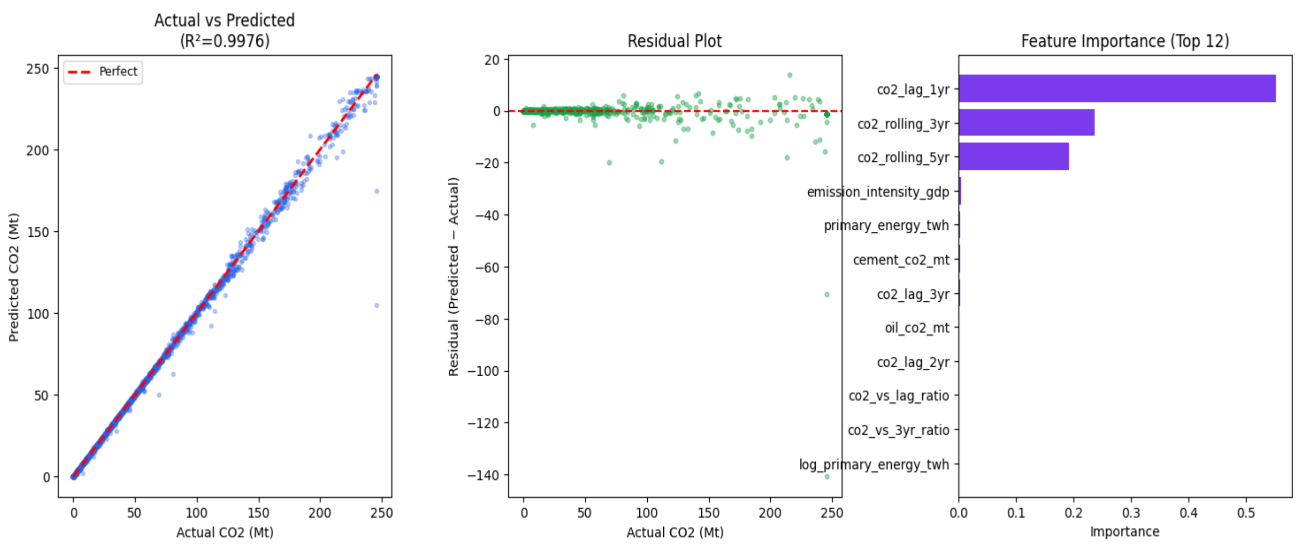

Figure 1: Model 1-emission predictor performance

Figure 2 Model 1 -Cross-validation and optimisation analysis. (Left) GroupKFold CV R^2 scores across 5 folds (mean=0.9983, std=0.0014), demonstrating stable generalisation across unseen countries. (Centre) Optuna hyperparameter optimisation history over 50 trials, showing convergence to best R^2=0.9993. (Right) Error distribution -median absolute percentage error of 1.7%, with most predictions within 10% of actual values. The XGBoost regressor is tuned using Optuna with 50 trials.

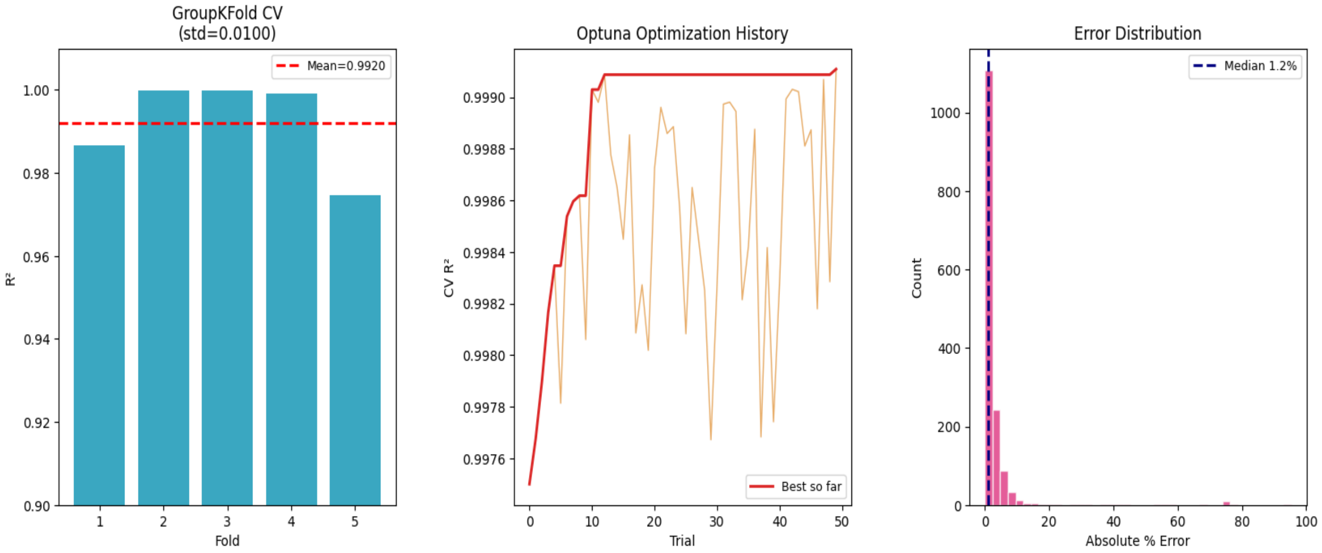

Figure 2: Model 1 – cross-validation and optimisation analysis

The best configuration identified at trial 32 uses 953 estimators, a maximum depth of 8, a learning rate of 0.1471, a subsample ratio of 0.5355, and a column sample ratio of 0.5884. The dataset is split 80-20 into training and test sets, yielding 6,198 training rows and 1,550 test rows, stratified by log-target quantile to ensure scale coverage in both splits. GroupKFold cross-validation with 5 folds by country is used throughout the tuning process to prevent temporal leakage across folds within-country.

3.4. Model 2: COPOD Anomaly Detector

The anomaly detection pipeline addresses the fundamental challenge that emission data spans vastly different scales across countries. A per-country z-score normalisation is applied first, where each feature value is expressed as the number of standard deviations from the respective country's historical mean. This transformation ensures that a doubling of emissions for a small Pacific island nation is equally detectable as an equivalent proportional anomaly in a major industrial economy, something that is impossible to achieve with global normalisation alone. Ground-truth labels for evaluation are generated using a rule-based protocol aligned with auditor expectations.

Type 1 anomalies are defined as a year-on-year change exceeding ±25 per cent. Type 2 anomalies are defined as coal fuel share shifts exceeding 20 percentage points. Type 3 anomalies are defined as a three-year rolling mean that rises while the global trend is declining. Approximately 10 percent of the dataset records carry an anomaly label under this scheme. Four candidate models are evaluated: Isolation Forest as a baseline, ECOD alone, an ECOD plus COPOD ensemble, and COPOD alone. COPOD is selected as the final model based on achieving the best AUC of 0.7152.

3.5. Model 3: XGBoost Trend Classifier

The trend classifier predicts a binary label: high_emission or low_emission. A country-year record is labelled high_emission if the entity's total CO_2 exceeds its own historical country median and the three-year rolling mean is simultaneously increasing. This relative formulation ensures that the classifier operates meaningfully at any emissions scale, whether for a large national economy or a small industrial facility. The feature set deliberately uses only ratio and percentage features rather than raw emission magnitudes. This design ensures scale invariance, meaning the model performs equally well whether applied to country-scale data from training or to company-level inputs from a small Indian SME. The final XGBoost classifier uses 300 estimators, balanced class weights to account for the approximately 28.6 per cent high_emission versus 71.4 per cent low_emission class imbalance, and a maximum depth of 8. Scale invariance is empirically validated: a country-scale input representing 150 Mt of coal yields a prediction confidence of 91.5%. In comparison, a realistic Indian SME input of 0.08 Mt of coal achieves 87.2% confidence in the same direction, confirming that the classifier generalises across scales.

3.6. CCTS Compliance Features

VayuCredit integrates four India-specific compliance features grounded in official BEE regulatory documents. The GEI formula is: GEI equals total tCO2e emitted divided by production output, in tonnes. Credits earned equal the product of (GEI target minus GEI actual) and production tonnes when the actual GEI is below the sector target. The penalty equals two times the average credit price multiplied by the shortfall tonnes when the actual GEI exceeds the target. The current carbon credit price used in calculations is between INR 830 and 1,000 per tonne CO2e, reflecting early 2026 India carbon market pricing. Non-compliant entities are required to pay penalties to CPCB within 90 days (Table 2).

Table 2: CCTS sector GEI targets (BEE, FY 2025-26)

Sector	GEI Targets	Compliance Year
Cement	0.054 tCO2e / tonne cement	2025-26
Aluminium	14.5 tCO2e / tonne Al	2025-26
Chlor-Alkali	1.20 tCO2e / tonne product	2025-26
Pulp and Paper	1.80 tCO2e / tonne paper	2025-26
Iron and Steel	2.20 tCO2e / tonne steel	2025-26
Petroleum Refinery	0.15 tCO2e / tonne output	Phase 2

The system architecture follows a three-layer design. The input layer collects company activity data through a structured form, which is validated using Pydantic v2 schemas with explicit input bounds before being passed to the ML layer. The ML layer consists of three trained models, serialised with joblib and loaded at application startup (Table 3).

Table 3: India-specific emission factors

Sector	Factor	Unit	Reference
Electricity	0.82	kg CO2/kWh	CEA India
Coal	2,618.0	kg CO2/tonne	BEE India (High Ash Content)
Diesel	2.68	kg CO2/litre	IPCC [1] Tier 1
Natural Gas	1.96	kg CO2/m3	IPCC [1] Tier 1
Cement	530.0	kg CO2/tonne	BEE India
Transport	0.21	kg CO2/km	IEA

The output layer returns predicted CO2 and percentile rank from Model 1, an anomaly flag and severity level from Model 2, a trend label with a CCTS-aligned recommendation from Model 3, along with a full GEI score and a credit or penalty amount in rupees. The backend is implemented in FastAPI with five REST endpoints: /api/calculate, /api/predict, /api/anomaly, /api/recommend, and /api/ccts.

4. Results and Discussion

All experiments were conducted in Python 3.10 using scikit-learn 1.6, XGBoost 2.0, PyOD 2.0, and optuma 3.6. THe training and evaluation were performed on standard hardware, demonstrating that the proposed framework does not require GPU resources for practical deployment.

4.1. Model 1: XGBoost Emission Predictor Results

Table 4 presents the evaluation metrics for the emission predictor, computed on the original emission scale after inverse-log transformation. The model achieves an R-squared of 0.9923 and a SMAPE of 4.40 percent, substantially exceeding the target thresholds of 0.87 and 28 percent respectively. The SMAPE metric is preferred over the standard MAPE because the latter inflates dramatically when actual values are near zero, as is common for small island nations with emissions below 0.1 Mt.

Table 4: Model 1 -XGBoost emission predictor results

Metric	Value
R-squared (R^2)	0.9923
Adjusted R-squared	0.9922
RMSE	7.65 Mt CO_2

MAE	0.76 Mt CO_2
SMAPE	4.40%
MAPE (filtered, emissions ≥ 1 Mt)	1.70%
CV R^2 (GroupKFold, 5-fold)	0.9983 ± 0.0014
CV Fold Score	[0.9972, 0.9995, 0.9998, 0.9989, 0.9962]
Train Rows	6,198
Test Rows	1,550

The filtered MAPE of 1.70 per cent, computed only for countries with emissions exceeding 1 Mt, indicates very strong performance among industrially relevant entities. The GroupKFold cross-validation score of 0.9983 ± 0.0014 confirms that the result is consistent and not driven by a favourable test split. The negative R-squared values for linear and ridge regression are expected and not indicative of implementation error. With a long-transformed target spanning 0.0001 to 245.7 Mt, linear models are mathematically unsuitable; a negative R-squared indicates the model performs worse than predicting the global mean for each observation (Table 5).

Table 5: Baseline model comparison (Emission Predictor)

Model	R2	RMSE(Mt)	MAE(Mt)
Linear Regression	-552,456,110	2,054,718	52,198
Ridge Regression	-578,975,773	2,103,457	53,436
Random Forest	0.9968	4.94	0.44
XGBoost	0.9982	3.73	0.49
XGboost with optuma	0.9923	7.65	0.76

The XGBoost model with Optuna tuning slightly trades raw RMSE for broader generalisation, which is appropriate for a model expected to handle inputs from diverse industrial contexts.

4.2. Model 2: COPOD Anomaly Detection Results

Table 6 presents the anomaly detection results across all four evaluated models. COPOD achieves the highest AUC (0.7152), the best F1 score (0.4817), and the best precision (0.4268) among all candidates. The improvement over the Isolation Forest baseline is 2.2 percent in AUC, which is meaningful for an unsupervised detection task on real-world data.

Table 6: Model 2 anomaly detection results

Model	F1 Score	Precision	Recall	AUC
Isolation Forest	0.4714	0.4151	0.5454	0.7052
ECOD	0.4715	0.4112	0.5524	0.7504
ECOD+ COPOD Ensemble	0.4811	0.3979	0.6083	0.7124
COPOD	0.4817	0.4268	0.5529	0.7152

For unsupervised anomaly detection on real, multi-country emission data, an F1 score of 0.40-0.55 is considered strong, given that the ground truth is inherently uncertain and the label distribution is approximately 10% anomalies. AUC is the primary reliability metric, and a value of 0.7152 indicates that the model meaningfully separates anomalous from normal years. The per-country z-score normalisation is one of the main contributions: without it, a doubling of emissions for a small nation such as Tuvalu from 0.005 to 0.01 Mt would be entirely invisible against the backdrop of absolute values dominated by large emitters.

4.3. Model 3: XGBoost Trend Classifier Results

Table 7 presents the comparison of all three classifier variants evaluated. The XGBoost classifier achieves the highest accuracy of 89.81 percent, a weighted F1 of 0.8968, and an AUC of 0.9562, and is selected as the final model.

Table 7: Model 3 classifier comparison

Model	Accuracy	F1(Weighted)	AUC	CV Accuracy
Random Forest	0.8910	0.8913	0.9482	0.85790
Gradient Boosting	0.89839	0.8819	0.9471	0.8547
XGBoost (Selected)	0.8981	0.8968	0.9562	0.8543

Table 8 provides a detailed per-class classification report. The high-emission recall is 0.78, meaning the model correctly identifies 78% of genuinely high-emission periods. The low-emission class achieves a recall of 0.94, indicating stronger support and a clearer distributional signal.

Table 8: XGBoost classifier — detailed classification report

Class	Precision	Recall	F1	Support
High Emission	0.85	0.78	0.81	443
Low Emission	0.92	0.94	0.93	1,107
Macro Average	0.88	0.86	0.87	1,550
Weighted Average	0.90	0.90	0.90	1,550

The scale invariance validation confirms that the model is not implicitly conditioned on country-scale magnitudes: the near-identical confidence values for country-scale and SME-scale inputs demonstrate that the ratio-only feature design achieves its intended purpose.

4.4. Key Contributions Summary

The six principal contributions of this work are as follows. First, the India-first design uses emission factors, GEI targets, and credit prices that are specific to India rather than global averages, making the platform directly actionable for regulated Indian entities. Second, the scale-invariant classifier design helps address the fundamental mismatch between country-level training data and company-level inference inputs. Third, per-country normalisation for anomaly detection ensures that small economies and large industrial nations receive equivalent detection sensitivity. Fourth, real CCTS integration uses actual BEE GEI targets from official notifications rather than hypothetical values. Fifth, one consistent unit system from kilograms to tonnes to megatonnes is maintained throughout, eliminating a common source of error in similar systems. Sixth, and last one is the three complementary models that cover prediction, monitoring, and recommendation in a unified framework.

5. Conclusion

This paper presented VayuCredit, an integrated AI-powered platform for carbon emission monitoring and regulatory compliance, developed specifically for Indian industries operating under the emerging Carbon Credit Trading Scheme. The platform addresses a critical gap which exists in carbon management tools. Most of these tools are either built for Western regulatory frameworks, such as the EU ETS, or rely on simplified rule-based calculations that don't account for the scale heterogeneity and sectoral diversity of Indian industrial operations. Three machine learning models were developed and validated on the OWID Global Carbon Project dataset of 7,748 records across 235 countries from 1990 to 2022. The emission predictor, an XGBoost regressor with log-transformed target and Optuna hyperparameter tuning, achieved an R^2 of 0.9923 and an SMAPE of 4.40 per cent, significantly outperforming linear and default tree-based baselines. The anomaly detector employs COPOD with per-country z-score normalisation. This design choice proved essential for detecting meaningful emission anomalies across countries whose absolute emissions differ by up to 7 orders of magnitude. The AUC of 0.7152 represents genuine discriminative ability rather than threshold-dependent accuracy, which is the appropriate evaluation criterion for unsupervised detection on real-world data. The trend classifier uses exclusively scale-invariant ratio features, which solves a fundamental problem encountered during development. A model trained on country-scale emission data (where China emits 12,000 Mt and Tuvalu emits 0.0001 Mt) initially produced only 50% confidence in realistic company-scale inputs.

By redesigning the feature set to include only fuel-mix percentages, year-over-year ratios, and rolling mean ratios, the classifier now produces confidence scores of 87–91 per cent at both the country and company scales, making it genuinely deployable for industrial use. Perhaps the most practically significant contribution of this work is the CCTS compliance engine. Rather than using hypothetical or placeholder regulatory targets, the system implements the actual Greenhouse Gas Emission Intensity targets notified by MoEFCC in October 2025 and January 2026, covering seven sectors: cement, aluminium, chlor-alkali, pulp and paper, petroleum refining, petrochemicals, and textiles. The GEI formula, credit calculation methodology, and penalty structure are all derived directly from the BEE Detailed Procedure for Compliance Mechanism under CCTS (July 2024 version). This means that when a cement plant enters its coal consumption, electricity usage, and production output into VayuCredit, the compliance score it receives is grounded in the same regulatory framework that will determine its actual financial obligations from FY 2025-26 onward. This level of regulatory fidelity distinguishes VayuCredit from existing academic carbon monitoring tools, which typically use simplified or generic compliance thresholds. The platform is delivered via a FastAPI backend that exposes 5 REST endpoints, with all 3 models loaded at startup and serving predictions in real time. The unit convention kg internally, displayed in tonnes and megatonnes with explicit labelling, was designed to eliminate the class of silent scientific errors that frequently occur in emission calculation systems, where megatonnes and gigatonnes are confused by ambiguous variable naming.

5.1. Future Work

While the current implementation establishes a strong foundation for AI-driven carbon compliance in Indian industries, several important directions remain for future development. The first and most immediate extension is integration of real-time data streams. Currently, VayuCredit operates on manually entered activity data. A company inputs its monthly coal consumption, electricity usage, and production volume, and the system calculates emissions and compliance scores for that reporting period. In practice, large industrial facilities increasingly deploy IoT-enabled energy meters, fuel flow sensors, and production counters that generate continuous data. Connecting VayuCredit to such data streams would transform it from a periodic reporting tool into a continuous monitoring system, enabling real-time anomaly alerts when, for example, a boiler malfunction causes an unexpected fuel consumption spike mid-month rather than discovering the anomaly only at the end of the compliance year. This is particularly relevant given that BEE requires annual monitoring reports under CCTS; early detection of compliance drift could allow facilities to take corrective action before the reporting deadline rather than facing penalties after the fact. The second direction involves extending the emission accounting scope from Scope 1 to full Scope 2 and Scope 3 coverage as defined by the GHG Protocol Corporate Standard.

The current platform focuses on direct combustion and process emissions (Scope 1) and provides basic coverage for purchased electricity (Scope 2). However, for many Indian industries, particularly in the textile and petrochemical sectors, Scope 3 emissions from raw material extraction, supplier logistics, and end-of-life product disposal account for a substantial share of the total carbon footprint. As global buyers increasingly require full value-chain emission disclosure (driven in part by the EU Carbon Border Adjustment Mechanism, which became fully operational in 2026), Indian exporters will face growing pressure to quantify and report Scope 3 emissions. Extending VayuCredit to support supplier emission surveys, logistics carbon calculation, and product lifecycle accounting would significantly increase its practical value for export-oriented manufacturers. The third direction is direct integration with India's carbon market infrastructure upon the commencement of CCC trading. The first compliance-based Carbon Credit Certificate trades are expected by mid-2026 through CERC-approved power exchanges, with the Grid Controller of India managing the central registry. At that point, VayuCredit could evolve from a compliance scoring tool into an active market participation platform that automatically calculates the number of tradeable CCCs a facility has earned, connects to the GCI registry API to verify credit issuance, and provides market timing recommendations based on projected credit price movements.

Given that credit prices are currently estimated at ₹830–1,000 per tonne CO_2e and the total CCTS-covered market represents approximately 700 million tonnes CO_2e annually, even a one-rupee improvement in trading timing across a portfolio of credits represents significant financial value. Beyond these three primary directions, researchers also see potential in extending the platform to support non-obligated entities through India's voluntary offset mechanism, which was operationalised in early 2026 and allows companies outside the nine mandatory CCTS sectors to earn credits through approved methodologies, including renewable energy, green hydrogen, and afforestation. Small and medium enterprises in sectors like sugarcane processing, food manufacturing, and general textiles, which currently fall outside mandatory CCTS coverage, could use VayuCredit to quantify their voluntary emission reductions and navigate the offset registration process, creating a pathway to carbon market participation that does not currently exist in an accessible, digitised form. VayuCredit represents an early yet substantive step toward making carbon-compliance infrastructure accessible, accurate, and actionable for the Indian industry at a moment when the regulatory stakes are becoming real. The work reported here was developed in parallel with the operationalisation of the CCTS itself, a coincidence of timing that researchers hope makes the contribution timely rather than merely theoretical.

Acknowledgement: The authors gratefully acknowledge the support and resources provided by SRM Institute of Science and Technology at Ramapuram, Dhaanish Ahmed College of Engineering, Technical University of Cotopaxi, and National Autonomous University of Mexico, which contributed significantly to the successful completion of this research.

Data Availability Statement: Data is available upon request from the corresponding author.

Funding Statement: This research received no financial support.

Conflicts of Interest Statement: No conflicts of interest were declared, and all references were appropriately cited.

Ethics and Consent Statement: The research adhered to ethical guidelines, with informed consent and confidentiality ensured.

References

1. IPCC, "Global Warming of 1.5°C: An IPCC Special Report," *Intergovernmental Panel on Climate Change*, 2018. [Accessed by 12/01/2025].

2. H. Ritchie, P. Rosado, and M. Roser, "CO2 and Greenhouse Gas Emissions," *Our World in Data*, 2023. [Accessed by 22/01/2025].
3. Government of India, "India's Updated First Nationally Determined Contribution Under Paris Agreement," *Ministry of Environment, Forest and Climate Change*, 2023. [Accessed by 25/01/2025].
4. International Carbon Action Partnership, "Indian Carbon Credit Trading Scheme," *ICAP*, 2025. [Accessed by 26/01/2026].
5. P. Friedlingstein, M. O'Sullivan, M. W. Jones, R. M. Andrew, J. Hauck, P. Landschützer, C. Le Quéré, H. Li, I. T. Luijkx, A. Olsen, G. P. Peters, W. Peters, J. Pongratz, C. Schwingshackl, S. Sitch, J. G. Canadell, P. Ciais, R. B. Jackson, S. R. Alin, A. Arneth, V. Arora, N. R. Bates, M. Becker, N. Bellouin, C. F. Berghoff, H. C. Bittig, L. Bopp, P. Cadule, K. Campbell, M. A. Chamberlain, N. Chandra, F. Chevallier, L. P. Chini, T. Colligan, J. Decayeux, L. M. Djeutchouang, X. Dou, C. Duran Rojas, K. Enyo, W. Evans, A. R. Fay, R. A. Feely, D. J. Ford, A. Foster, T. Gasser, M. Gehlen, T. Gkritzalis, G. Grassi, L. Gregor, N. Gruber, Ö. Gürses, I. Harris, M. Hefner, J. Heinke, G. C. Hurtt, Y. Iida, T. Ilyina, A. R. Jacobson, A. K. Jain, T. Jarníková, A. Jersild, F. Jiang, Z. Jin, E. Kato, R. F. Keeling, K. Klein Goldewijk, J. Knauer, J. I. Korsbakken, X. Lan, S. K. Lauvset, N. Lefèvre, Z. Liu, J. Liu, L. Ma, S. Maksyutov, G. Marland, N. Mayot, P. C. McGuire, N. Metzl, N. M. Monacci, E. J. Morgan, S. I. Nakaoka, C. Neill, Y. Niwa, T. Nützel, L. Olivier, T. Ono, P. I. Palmer, D. Pierrot, Z. Qin, L. Resplandy, A. Roobaert, T. M. Rosan, C. Rödenbeck, J. Schwinger, T. L. Smallman, S. M. Smith, R. Sospedra-Alfonso, T. Steinhoff, Q. Sun, A. J. Sutton, R. Séférian, S. Takao, H. Tatebe, H. Tian, B. Tilbrook, O. Torres, E. Tourigny, H. Tsujino, F. Tubiello, G. Van Der Werf, R. Wanninkhof, X. Wang, D. Yang, X. Yang, Z. Yu, W. Yuan, X. Yue, S. Zaehle, N. Zeng, and J. Zeng, "Global Carbon Budget 2024," *Earth System Science Data*, vol. 17, no. 3, pp. 965–1039, 2025.
6. T. Ahmad and H. Chen, "A review on machine learning forecasting growth trends and their real-time applications in different energy systems," *Sustainable Cities and Society*, vol. 54, no. 3, p. 102010, 2020.
7. L. Zhang, G. Lu, X. Yan, P. Xia, Z. Chen, and D. Wu, "A differential evolution optimized hybrid XGBoost for accurate carbon emission prediction," *Environmental Modelling & Software*, vol. 193, no. 9, p. 106627, 2025.
8. J. Liang and X. Liu, "High-Resolution Spatiotemporal Carbon Emission Estimation in Northeast China based on XGBoost and Multi-Source Data," *Applied Sciences*, vol. 16, no. 5, p. 2272, 2026.
9. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, San Francisco, California, United States of America, 2016.
10. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD '19),* Anchorage, Alaska, United States of America, 2019.
11. F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, 2008.
12. Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: Copula-based outlier detection," in *2020 IEEE Int. Conf. Data Mining (ICDM)*, Sorrento, Italy, 2020.
13. Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. H. Chen, "ECOD: Unsupervised outlier detection using empirical cumulative distribution functions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12181–12193, 2023.
14. Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.
15. G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 1, pp. 891–927, 2016.
16. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
17. C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, no. 4, pp. 70–83, 2018.
18. B. Li, S. Song, Y. Gu, H. Yu, Y. Chai, and F. Wu, "Carbon emission prediction model based on LSTM enhanced by attention mechanism and Elman neural network," in *2024 IEEE/ACIS 24th Int. Conf. Computer and Information Science (ICIS*), Shanghai, China, 2024.
19. A. Mardani, D. Štreimikienė, F. Cavallaro, N. Loganathan, and M. Khoshnoudi, "Carbon dioxide emissions and economic growth: A systematic review of two decades of research from 1995 to 2017," *Science of the Total Environment*, vol. 649, no. 2, pp. 31–49, 2019.
20. Ministry of Environment, Forest and Climate Change, "Greenhouse Gas Emission Intensity (GEI) targets," *Ministry of Environment, Forest and Climate Change*, 2026. [Accessed by 27/01/2026].
21. Ministry of Environment, Forest and Climate Change, "Government notifies Greenhouse Gas Emission Intensity Targets for 208 more carbon-intensive industries," *Press Information Bureau*, 2026. [Accessed by 27/01/2026].
22. Central Electricity Authority, "CO_2 Baseline Database for the Indian Power Sector," *Ministry of Power, Government of India*, 2025. [Accessed by 28/01/2026].
23. European Commission, "EU Emissions Trading System (EU ETS)," *Climate Action*, 2024. [Accessed by 28/01/2025].

24. Steadfast Consultants, "India's Carbon Trading Scheme: A Pioneering Step Towards Climate Action," *SBC*, 2025. [Accessed by 28/01/2026].
25. R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 3rd ed., *OTexts*, Waterloo, Canada, 2021.
26. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.