Machine Learning for Carbon Emissions Prediction and Anomaly Detection: A Comparative Study of Traditional and Deep Learning Approaches

Rafiu Mope ISIAKA[1], Akeem Femi KADRI[2], Oluwasayo EKUNDAYO[3], Habeeb Olayinka SULAIMAN[4]

1,2,3,4 Department of Computer Science, Kwara State University, Malete, Nigeria.

KWASU Journal of Information, Communication and Technology (KJICT.)2025, 4(1)
© 2025 Faculty of Information and Communication Technology, Kwara State University, Malete, Nigeria.
URL: http://kjict.ng/

Abstract

Accurate greenhouse gas emissions monitoring is critical for evaluating climate policies and ensuring regulatory compliance. Although machine learning offers promising capabilities, systematic comparisons between traditional and deep learning approaches for structured environmental time-series data remain limited. Here, we evaluated seven prediction models and three anomaly detection approaches on monthly CO_2 emissions data from 20 major emitting countries between 2000 and 2024. Random Forest achieved superior performance ($R^2 = 0.902$) compared to all deep learning architectures ($R^2 \approx 0$), while requiring 15–60 times less training time. SHapley Additive exPlanations revealed that the 12-month lag feature dominated the predictions (75.6% importance), indicating strong year-over-year persistence with direct policy implications. Deep learning failure stems from insufficient training data (3,863 samples), structured tabular characteristics with domain-engineered features, and unnecessary architectural complexity, thereby aligning with emerging evidence that traditional methods often outperform deep learning on tabular datasets. Anomaly detection proved challenging, with all unsupervised models achieving F1-scores below 0.05, suggesting that supervised approaches with labelled examples are necessary for operational deployment. We recommend prioritising Random Forest for prediction, investing in domain-informed feature engineering, integrating explainability tools for transparency, and deploying anomaly detection with human review. We identified eight research frontiers, including satellite data integration, federated learning, causal inference, and hybrid physics-machine learning models. This study demonstrates that traditional machine learning, combined with domain expertise and interpretability tools, can provide accurate, efficient, and transparent emission predictions for operational climate monitoring systems.

Keywords: Machine learning, emissions prediction, anomaly detection, interpretable AI, climate monitoring

1. Introduction

The escalating climate crisis demands unprecedented precision in monitoring and reducing greenhouse gas (GHG) emissions to achieve global climate targets. The World Meteorological Organisation confirmed 2024 as the warmest year on record, with global temperatures reaching 1.55°C above pre-industrial levels (WMO, 2025), perilously close to the 1.5°C threshold established by the Paris Agreement. This alarming trajectory underscores the critical need for enhanced monitoring and predictive capabilities to inform effective climate actions.

The Paris Agreement requires countries to limit global warming to well below 2°C, preferably to 1.5°C, through Nationally Determined Contributions (NDCs), necessitating that global GHG emissions peak before 2025 and decline by 43% by 2030 (UNFCCC, 2025). Traditional emissions monitoring systems rely on manual reporting, periodic sampling, facility-level continuous emissions monitoring systems (CEMS), and bottom-up inventory approaches (U.S. EPA, 2023). While foundational, these methods suffer from critical limitations, such as data quality challenges, including anomalies, incomplete records, and reporting inconsistencies (Osswald & MBA, 2025).

Other constraints include temporal and spatial resolution, where ground-based stations provide high-quality point measurements but limited spatial coverage, particularly in remote regions and developing countries (Nakalembe et al., 2024). Transparency and verification gaps exist, where the absence of independent verification mechanisms for self-reported data creates opportunities for greenwashing (Carbon Market Watch, 2024).

Recent advancements in machine learning (ML) offer transformative potential for overcoming these limitations. Ensemble methods, such as Random Forest and XGBoost, have achieved high accuracy in predicting GHG emissions across various sectors. Begum and Mobin (2025) achieved 96.21% prediction accuracy for CO_2 emissions for the 11 highest-emitting nations using six ML models with sequential feature selection. Bista et al. (2025) and Jin and Sharifi (2025) demonstrated that ML methods can effectively predict greenhouse gas emissions in urban environments for current and future scenarios.

Deep learning architectures, including LSTM networks, GRUs, and attention-based models, have shown promise in capturing complex temporal dependencies in emissions time series (Wang et al., 2023). Wang et al. (2025) developed a hybrid LSTM-GRU model for predicting annual carbon emissions across Omani industries, demonstrating that combining architectures can leverage the strengths of both approaches. CNNs have been applied to satellite imagery and gridded emissions data, whereas hybrid CNN-RNN architectures can simultaneously extract spatial features and model temporal dynamics (Yuan et al., 2025).

Anomaly detection techniques, including isolation forests, one-class SVMs, and autoencoders, can identify unusual emission patterns that deviate from the expected baselines (Yepmo et al., 2024). These capabilities enable the early detection of unexpected events, such as methane leaks from oil and gas infrastructure, industrial process upsets, and unreported combustion activities (Dayalu et al., 2024; Yazdinejad et al., 2025).

Interpretability tools, such as SHAP and LIME, have emerged to address "black box" concerns in high-stakes policy contexts. Latif et al. (2025) developed an interpretable ML framework for predicting urban air quality, demonstrating that SHAP values can identify the most influential factors driving predictions, thereby enabling targeted interventions. Bommer et al. (2024) provided guidance for evaluating and ranking XAI methods in climate science, emphasising that no single method is universally superior and that practitioners should consider multiple approaches.

Despite these advances, three critical gaps limit the operational deployment of ML-based emission monitoring systems. First, systematic comparisons between traditional ML and deep learning approaches for structured environmental time-series data are limited. While individual studies have demonstrated success with specific models, comprehensive benchmarking across multiple model families, such as ensemble methods, deep learning, and hybrid architectures on consistent datasets, is lacking. Recent research suggests that simpler models can sometimes outperform deep learning, particularly with limited datasets or when interpretability is prioritised (MIT, 2025). However, empirical evidence in the emissions domain remains insufficient. Second, the relative importance of model interpretability versus predictive accuracy in operational contexts is poorly understood in the literature. Although SHAP and LIME have been applied to environmental monitoring, their specific application to emissions prediction, particularly for informing policy interventions, remains underexplored. The policy relevance of feature importance patterns, such as lag-dominant emissions behaviour, has not been explicitly connected to climate action strategies. Finally, anomaly detection in emissions monitoring faces fundamental challenges. Existing studies have primarily focused on prediction tasks, with limited evaluation of unsupervised anomaly detection methods for identifying irregularities, equipment failures, or non-compliance events. The operational viability of unsupervised and supervised approaches for real-world deployment remains unclear.

To address these gaps, this study makes three contributions. First, it provides a systematic empirical comparison of seven prediction models (LSTM, GRU, Attention-based BiGRU, CNN-RNN hybrid) and three anomaly detection approaches (Isolation Forest, One-Class SVM, Autoencoder) on monthly CO_2 emissions data from 20 major emitting countries (2000-2024). This comprehensive benchmarking of Random Forest, XGBoost, Support Vector Regression quantifies the performance trade-offs between traditional ML and deep learning for structured emissions data, providing evidence-based guidance for model selection. Second, we conducted an in-depth interpret-

ability analysis using SHAP to identify the key drivers of emissions predictions and explicitly connect these findings to climate policy implications. By quantifying feature importance patterns (particularly lag-dominant behaviour), we demonstrate how interpretable ML can inform policy design, including the identification of emission inertia and evaluation of intervention timing. Third, this study evaluated the anomaly detection viability for operational deployment, assessed the performance of unsupervised methods on synthetic emission anomalies, and provided practical recommendations for human-in-the-loop systems. This addresses the critical need for transparent and explainable AI systems in environmental monitoring contexts, where regulatory compliance and stakeholder trust are paramount.

This study leverages publicly available emissions data from 'Our World in Data', supplemented with synthetically generated monthly and sector-specific data to simulate real-world monitoring scenarios (Ritchie et al., 2024). A rigorous multiphase methodology encompasses data preprocessing and feature engineering, model development and architecture design, time-aware cross-validation, comprehensive performance evaluation, interpretability analysis, and conceptual MLOps deployment pipeline design with drift detection capabilities. By systematically addressing these gaps, this study provides methodological insights and practical tools for achieving global climate targets, enabling policymakers, industries, and communities to better track, understand, and mitigate greenhouse gas emissions through evidence-based, interpretable, and operationally viable ML systems.

2. Literature Review

The application of machine learning to emissions monitoring has witnessed remarkable growth, driven by expanding datasets, increasing computational capabilities, and urgent climate action requirements. This section reviews recent advancements in emissions prediction, anomaly detection, and model interpretability, focusing on the three research gaps identified in the Background: systematic model comparisons, interpretability-accuracy trade-offs, and anomaly detection viability.

Traditional machine learning algorithms, particularly ensemble methods, have demonstrated exceptional performance in emission-forecasting tasks. Random Forest and XGBoost have emerged as dominant approaches because of their ability to capture nonlinear relationships while maintaining computational efficiency and interpretability (Rahman et al., 2023).

Begum and Mobin (2025) developed a comprehensive machine learning pipeline for predicting CO_2 emissions for the 11 highest-emitting nations, achieving 96.21% prediction accuracy. Their study employed six ML models with sequential feature selection, incorporating 11 economic, industrial, and energy consumption variables, demonstrating that ensemble methods can effectively integrate heterogeneous data sources to produce policy-relevant forecasts aligned with the Paris Agreement targets. However, their study did not evaluate model explainability or compare the performance with deep learning approaches.

Khajavi and Rastgoo (2023) implemented a hybrid approach combining Random Forest, Support Vector Regression, and Response Surface Methodology to predict CO_2 emissions in 30 major Chinese cities. Their work highlighted the importance of spatial heterogeneity in emission patterns and the need for city-specific calibration but similarly lacked systematic comparison with deep learning methods or interpretability analysis (Ene Yalçın, 2024)

Yalçın (2024) developed a forecasting framework based on advanced ML algorithms for GHG emissions, demonstrating that ensemble methods can achieve high accuracy across diverse sectors. These studies collectively establish ensemble methods as effective for emissions prediction but leave open questions regarding their relative performance compared to deep learning and the policy relevance of their predictions.

Deep learning approaches have shown promise for capturing complex temporal dependencies in emissions time series. Wang et al. (2025) developed a hybrid LSTM-GRU model for predicting annual carbon emissions across Omani industries, demonstrating that combining architectures can leverage LSTM's capacity for long-term memory and GRU's computational efficiency. The model effectively captures complex temporal dependencies and nonlinear patterns, providing industry-specific reduction pathways that align with national climate commitments.

Yuan et al. (2025) applied attention-based CNN-LSTM architectures to atmospheric ozone prediction, demonstrating that hybrid models can simultaneously extract spatial features and model the temporal

dynamics. Their work showed that attention mechanisms can improve the model's focus on relevant time steps, enhancing prediction accuracy.

However, recent studies have revealed the limitations of deep learning approaches. Derot et al. (2024) found that applying signature methods could enhance the predictive performance of both traditional and deep learning models, suggesting that data preprocessing may be more important than model complexity. Critically, a 2025 MIT study demonstrated that simpler models can sometimes outperform deep learning in climate prediction tasks, particularly when datasets are limited or interpretability is prioritised (MIT, 2025). This finding underscores the importance of matching model complexity to data characteristics and application requirements, which is a gap that this study addresses through a systematic comparison.

The growing emphasis on explainable artificial intelligence (XAI) in environmental applications reflects the need for transparent and trustworthy models that can inform policy decisions. SHAP and LIME are standard tools for interpreting complex machine learning models (Lundberg & Lee, 2017). Latif et al. (2025) developed an interpretable ML framework for predicting urban air quality, demonstrating that SHAP values can identify the most influential factors driving predictions, thereby enabling targeted interventions. Their study emphasised that policymakers can be optimistic about applying interpretable ML frameworks in environmental monitoring systems but did not explicitly connect the feature importance patterns to specific policy recommendations.

Bommer et al. (2024) provided comprehensive guidance for evaluating and ranking XAI methods in climate science applications. Their review emphasised that the selection of appropriate explainability techniques depends on the specific use case, stakeholder needs, and model architecture, highlighting that no single XAI method is universally superior to others. They recommended that practitioners consider multiple approaches to gain a comprehensive understanding.

Despite these advances, the policy relevance of interpretability remains underexplored. While SHAP and LIME can identify important features, translating these technical explanations into actionable policy recommendations, such as understanding emission inertia or optimal intervention timing, requires explicit analysis that existing studies have not provided. This study addresses this gap by linking SHAP-derived feature importance patterns to climate policy implications.

Anomaly detection is increasingly important for identifying unusual emission events, equipment malfunctions, and potential noncompliance with environmental regulations. Unsupervised and semi-supervised learning approaches show promise for detecting novel or rare events without labelled training data.

Yepmo et al. (2024) leveraged Isolation Forest for anomaly detection, demonstrating that it can effectively identify outliers in high-dimensional emissions data without requiring labelled training examples. Their research aimed to extract contrastive explanations between anomalies and the intrinsic structure of regular data points.

Autoencoder-based approaches have gained traction because of their ability to learn compressed representations of normal emission patterns and identify deviations. Üstek et al. (2024) proposed deep autoencoders for unsupervised anomaly detection in wildfire prediction, demonstrating that autoencoders can learn latent representations of environmental conditions and flag anomalous patterns indicative of elevated risk.

Dayalu et al. (2024) applied Isolation Forest for PM2.5 anomaly detection in air quality monitoring, achieving promising results for identifying exceptional pollution events. Yazdinejad et al. (2025) demonstrated improved spatial resolution and anomaly classification in methane emission detection using AI-driven approaches, showing that deep learning can enhance satellite-based anomaly detection.

However, systematic evaluations of the anomaly detection performance for emissions monitoring remain limited. Existing studies primarily focus on prediction tasks, with limited assessments of precision, recall, and F1-scores for anomaly detection. The operational viability of unsupervised versus supervised approaches, particularly for regulatory compliance applications that require high precision, remains unclear. This study addresses this gap through a comprehensive evaluation of three anomaly detection methods with quantified performance metrics.

Few studies have systematically compared traditional ML and deep learning approaches for emission prediction using consistent datasets. Most research focus on demonstrating success with a single model family rather than providing empirical evidence for model-selection decisions.

Zabërgja et al. (2025) conducted a comprehensive benchmark comparing deep learning and decision trees on tabular data, finding that simpler variants of traditional MLP architectures often match or exceed the performance of complex deep learning pipelines. Hollmann et al. (2025) demonstrated that deep learning methods traditionally struggle with tabular data because of the heterogeneity between datasets and features within datasets.

These findings from general machine learning research suggest that emissions data are fundamentally tabular with well-defined features that may favour traditional ML methods. However, empirical evidence in the emissions domain is lacking. This study fills this gap by systematically comparing seven prediction models (three traditional ML and four deep models) on structured emissions time-series data.

Despite substantial progress, three critical gaps remain: systematic model comparison, policy-relevant interpretability, and anomaly detection viability. For systematic model comparison, existing studies have demonstrated success with individual models but lack comprehensive benchmarking across traditional ML and deep learning on consistent emissions datasets. However, the relative performance, computational efficiency, and interpretability trade-offs of these methods remain unclear. Regarding policy-relevant interpretability, although SHAP and LIME have been applied to environmental monitoring, explicit connections between feature importance patterns, such as lag-dominant behaviour, and climate policy implications, such as emission inertia and intervention timing, are absent. The operational value of interpretability for policymakers has not been thoroughly explored. Finally, for anomaly detection viability, systematic evaluation of anomaly detection methods for emissions monitoring with quantified performance metrics, such as precision, recall, and F1-scores, is limited. The operational feasibility of unsupervised approaches for regulatory compliance applications that require high precision is unclear.

This study addresses these gaps by systematically comparing seven prediction models and three anomaly detection approaches using monthly CO_2 emissions data from 20 major emitting countries, providing empirical evidence for model selection. In addition, by conducting an in-depth SHAP analysis with explicit connections to climate policy implications, we demonstrate how interpretable ML can be used to inform policy design. Finally, we evaluated the anomaly detection viability using comprehensive performance metrics and practical recommendations for operational deployment. By filling these gaps, this study contributes to evidence-based emission monitoring and climate action.

3. Material and Method

This study adopted a systematic multiphase approach to develop, evaluate, and deploy machine learning models for emissions prediction and anomaly detection. The methodology encompasses six phases: data collection and preprocessing, model development, training and validation, interpretability analysis, deployment framework design, and comprehensive evaluation.

Data Collection and Preprocessing

The primary data source was the Our World in Data (OWID) CO_2 and Greenhouse Gas Emissions dataset (Ritchie et al., 2024), which aggregates emissions data from authoritative sources, including the Global Carbon Project, BP Statistical Review, and national environmental agencies. Data were filtered to focus on 20 major emitting countries (USA, China, India, Germany, UK, Japan, Brazil, Canada, Australia, France, Italy, Spain, South Korea, Mexico, Indonesia, Saudi Arabia, Russia, South Africa, Turkey, and Poland) over 2000-2024, yielding 480 annual records representing approximately 75% of global CO_2 emissions.

To enable granular temporal analysis suitable for time-series models, synthetic monthly emissions data were generated from annual records by incorporating seasonal patterns, higher winter emissions due to heating demand, and stochastic noise to simulate realistic intra-annual variability. This resulted in 5,760 monthly records (20 countries × 24 years × 12 months). Additionally, synthetic sector-specific data were created for seven sectors (Power Generation, Transportation, Industry, Residential, Agriculture, Waste Management, Land Use), totalling 3,360 records.

While real monthly emissions data are ideal, such granular temporal resolution is unavailable for most countries in historical datasets. Synthetic monthly data generation is a standard approach in emissions modelling research (Begum & Mobin, 2025) and enables evaluation of model capabilities for high-frequency monitoring scenarios increasingly relevant with satellite-based systems. The seasonal patterns and stochastic noise were calibrated to reflect known

emission dynamics, and model validation focused on relative performance comparisons rather than absolute prediction accuracy.

Standard preprocessing procedures were applied, including missing value imputation using hierarchical forward-fill, backward-fill, and median imputation. Also, temporal alignment to consistent monthly resolution (YYYY-MM-01) format, and feature scaling using StandardScaler zero mean, and unit variance. Similarly, to evaluate anomaly detection models, synthetic anomalies were injected into 2% of records (115 anomalies) by multiplying normal values by factors of 1.5-3.0 (spikes) or 0.3-0.6 (drops).

Effective feature engineering is crucial for capturing temporal dependencies. The engineered feature set comprised the 1, 3, 6, and 12-month lags to capture autoregressive patterns, also the 3, 6, and 12-month rolling averages to smooth fluctuations and Month-over-month (MoM) and year-over-year (YoY) growth rates. To capture seasonal patterns, cyclical time features were constructed using sine and cosine transformations of the monthly variables: month_sin = sin(2π × month / 12) and month_cos = cos(2π × month / 12). Additionally, discrete month (1-12) and quarter (1-4) features were included.

The final feature set comprised 13 engineered features plus the target variable (monthly CO_2 emissions). A time-aware sequential split 70:15:15 for training, validation and testing respectfully preserved temporal integrity, yielding 3,863, 829, and 828 records respectively, ensuring models trained on historical data were evaluated on future periods.

Model Development and Architecture Design

Seven diverse prediction models were implemented to provide comprehensive comparison of traditional ML and deep learning approaches. The traditional ML models are Random Forest (RF), XGBoost and Support Vector Regression (SVR). Also, the Deep learning models used include Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Attention-based Bidirectional GRU and CNN-RNN Hybrid model. A brief description of the models, configuration, hyperparameters, and training process, as applicable are as follows.

The Random Forest Regressor is an ensemble method that constructs multiple decision trees during training and outputs the mean prediction. The hyperparameter settings included 100 estimators, maximum depth: none, minimum samples split: 2, minimum samples leaf: 1, random state: 42, and number of jobs: -1.

The XGBoost Regressor is a gradient boosting framework that builds an ensemble of weak learners sequentially, with each tree correcting the errors of its predecessors. The hyperparameters set for the model include 100 estimators; 0.1 learning rate; maximum depth: 6; subsample: 0.8; column sample by tree: 0.8; random state: 42; and number of jobs: -1.

Support Vector Regression (SVR) is a kernel-based method that finds a hyperplane in a high-dimensional feature space to minimise the prediction error. Owing to computational constraints with large datasets, the SVR was trained on a subset of 30% of the training data (1,158 samples). The hyperparameters are Kernel: Radial Basis Function (RBF); C (regularisation parameter): 1.0; Epsilon (tube width): 0.1; Gamma: 'scale' (1 / (n_features × X.var())).

Long Short-Term Memory (LSTM) is a recurrent neural network architecture designed to capture long-term dependencies in sequential data using gating mechanisms. The LSTM model architecture consists of the following: Input layer: Shape (12, 13) representing 12 time steps with 13 features each; LSTM layer 1:64 units with ReLU activation and return sequences; Dropout layer 1:20% dropout rate for regularisation; LSTM layer 2:32 units with ReLU activation; Dropout layer 2:20% dropout rate; Dense output layer: 1 unit with linear activation

The Gated Recurrent Unit (GRU) is a simplified variant of LSTM with fewer parameters, using reset and update gates instead of separate input, forget, and output gates. The architecture was as follows: input layer: shape (12, 13); GRU layer 1:64 units with ReLU activation and return sequences; dropout layer 1:20% dropout rate; GRU layer 2:32 units with ReLU activation; dropout layer 2:20% dropout rate; dense output layer: 1 unit with linear activation.

The Attention-based Bidirectional GRU is a hybrid architecture that combines bidirectional GRU layers (processing sequences in both forward and backward directions) with an attention mechanism that learns to weight the importance of different time steps. The architecture comprises an input layer: shape (12, 13); Bidirectional GRU layer: 64 units (32 forward + 32 backward) with tanh activation; Custom Attention layer: learns attention weights over time steps; and dense output layer: 1 unit with linear activation.

The CNN-RNN Hybrid model uses convolutional layers to extract local temporal features, followed by recurrent layers to model long-term dependencies. The architecture consists of an input layer: shape (12, 13); Conv1D layer: 64 filters, kernel size 3, ReLU activation; MaxPooling1D layer: pool size 2; GRU layer: 32 units with ReLU activation; and dense output layer: 1 unit with linear activation.

All deep learning models used the Adam optimizer (learning_rate=0.001), MSE loss, batch_size=32, and early stopping (patience=5 epochs monitoring validation loss). Additionally, the LSTM used 50 maximum epochs, and the GRU used 26 epochs trained before early stopping. For the Bidirectional GRU and CNN-RNN Hybrid, 10 epochs were trained using the Adam optimizer and MSE loss.

The following are descriptions of the three unsupervised and semi-supervised anomaly detection models implemented. The isolation forest is an ensemble method that isolates anomalies by randomly partitioning the feature space. The hyperparameters were as follows: number of estimators, 100; contamination, 0.02; random state, 42; and number of jobs, -1. The one-class SVM is a support vector machine variant trained to learn a decision boundary around normal data points and classify points far from this boundary as anomalies. Owing to computational constraints, the model was trained using 30% of the training data. The hyperparameters included Kernel: RBF; Nu: 0.02 (upper bound on fraction of outliers); Gamma: 'scale.' An autoencoder is a neural network trained to reconstruct input data using a compressed bottleneck layer. Anomalies were detected based on a high reconstruction error. The Architecture of the encoder and decoder is as follows: Encoder: Dense layer 1:32 units with ReLU activation. Dense layer 2:16 units with ReLU activation; bottleneck layer: eight units with ReLU activation. The decoder has Dense layer 1:16 units with ReLU activation; Dense layer 2:32 units with ReLU activation; and output layer: 13 units with linear activation (reconstructed features). The training configuration included an optimizer: Adam with a learning rate of 0.001; loss function, MSE; Batch size, 32; epochs, 50; and anomaly threshold, 95th percentile of reconstruction errors on training data.

Training and Validation

All models were trained on 3,863 samples using an engineered feature set. Deep learning models used 12-step sliding windows as inputs, with the 13th step as the target, maximising temporal data utilisation. Early stopping prevented overfitting by terminating the training if the validation loss did not improve for 5 consecutive epochs.

The components of the computational environment included Ubuntu 22.04, Python 3.11, scikit-learn 1.5.2, XGBoost 2.1.3, TensorFlow 2.18.0, Keras 3.7.0. Hyperparameters were selected based on established best practices and preliminary experiments. However, a comprehensive grid search was not performed because of computational constraints.

Evaluation Metrics

The prediction models were evaluated using the following three complementary regression metrics: The Root Mean Squared Error (RMSE), which measures the average magnitude of the prediction errors and penalise larger errors; the Mean Absolute Error (MAE), which provides a linear measure of the average prediction error that is more robust to outliers than the RMSE; and the coefficient of determination (R^2), which quantifies the proportion of variance in the target variable explained by the model, ranging from negative infinity (worse than predicting the mean) to 1 (perfect predictions).

Similarly, the anomaly detection models were evaluated using classification metrics adapted for imbalanced data, including precision, recall, and F1 score. Precision measures the proportion of detected anomalies that are real. Recall measures the proportion of actual anomalies that are successfully detected. The F1-score provides a harmonic mean of precision and recall, balancing both metrics.

Model Interpretability and Explainability

Feature importance was extracted from tree-based models (Random Forest using the mean decrease in impurity; XGBoost using the gain metric). SHAP (SHapley Additive exPlanations) provided model-agnostic interpretability for the best-performing model (Random Forest), quantifying the marginal contribution of each feature to predictions based on cooperative game theory (Lundberg & Lee, 2017). SHAP values were computed for 200 random test instances using TreeExplainer, with summary and bar plots visualising feature contributions and impact directions.

Deployment and Monitoring Framework

A conceptual MLOps deployment pipeline was designed to demonstrate operational implementation

pathways, including the Model Registry, Drift Detection, RESTful API, Monitoring Dashboard, and Reproducibility.

The Model Registry tracks trained models with metadata, including file paths, serialisation formats, performance metrics, training timestamps, version numbers, and deployment statuses. The Drift Detection Framework monitors data distribution changes and performance degradation using methods (Kolmogorov-Smirnov test, Population Stability Index (PSI), Jensen-Shannon Divergence), alert thresholds (feature drift (0.05), prediction drift (0.10), performance degradation (0.15)), and monitoring frequency (daily comparison of 7-day recent window against 30-day historical reference)).

A RESTful API was implemented using FastAPI to enable real-time predictions. The API endpoints include: /predict: Accepts feature vectors and returns emission predictions and anomaly flags; /health: Returns system health status; /models: Lists available models and their versions and /features: Returns required feature specifications

Furthermore, a conceptual monitoring dashboard was designed with six panels: 1. Prediction accuracy over time (RMSE, MAE, R^2), 2. Anomaly detection rate (gauge visualisation), 3. Feature drift detection (heatmap), 4. Model latency (histogram) 5. Request volume (counter) and 6. Error rate (line chart). Three alert types were configured: high error rate (>5%), performance degradation ($R^2 < 0.85$), and feature drift (drift score > 0.10).

To ensure reproducibility, all random number generators were seeded with a fixed value (42) across all models and data processing steps. The complete code, trained models, and processed datasets are available in a project repository. The experimental pipeline was designed to be executed sequentially using six Python scripts, each corresponding to a phase of this methodology.

4. Result and Discussion

This section presents a comprehensive evaluation of the seven prediction and three anomaly detection models developed in this study. The analysis focuses on the comparative performance across multiple metrics, computational efficiency, model interpretability, and practical implications for the operational deployment. All results are reported on the held-out test set to provide unbiased estimates of the generalisation performance.

Prediction Model Performance

The results of the prediction capabilities of the developed models for the test data are presented in the following five subsections.

Overall Performance Comparisons

Table 1 presents the performance of all seven prediction models on the test set, evaluated using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The training time was also reported to evaluate the computational efficiency.

Table 1: Prediction Model Performance on Test Set

Model	Test RMSE	Test MAE	Test R^2	Training Time (s)	Epochs Trained
Random Forest	71.74	10.04	0.9021	0.71	N/A
XGBoost	92.46	13.98	0.8374	0.09	N/A
SVR	103.86	18.27	0.7948	0.23	N/A
LSTM	227.50	125.18	0.0106	25.07	16
GRU	227.96	121.21	0.0065	42.57	26
Attention BiGRU	229.64	115.07	-0.0082	16.74	10
CNN-RNN	230.63	108.64	-0.0169	11.86	10

The results revealed a remarkable performance hierarchy, with traditional machine learning models substantially outperforming deep learning architectures. Random Forest achieved the best overall performance, with an R^2 of 0.9021, explaining approximately 90% of the variance in monthly CO_2 emissions. This corresponds to an RMSE of 71.74 million tonnes and an MAE of 10.04 million tonnes, indicating high predictive accuracy across the range of emission values in the test set.

XGBoost ranked second with an R^2 of 0.8374,

demonstrating a strong performance, while being the most computationally efficient model with a training time of only 0.09 s. The gradient boosting approach effectively captured the non-linear relationships in the data, although it fell short of the Random Forest ensemble averaging benefits.

Support Vector Regression (SVR) achieved an R^2 of 0.7948, placing it third among all models. Although respectable, SVR's performance of SVR was constrained by training on only 30% of the available data owing to computational limitations. The RBF kernel successfully captured nonlinear patterns. However, the reduced training set size likely limited its ability to learn the full complexity of emission dynamics.

The deep learning models exhibited poor performance, with all four architectures (LSTM, GRU, Attention BiGRU, and CNN-RNN) achieving R^2 values near or below zero. The LSTM model achieved the highest R^2 among the deep learning approaches at 0.0106, indicating that it barely outperformed a naive baseline that simply predicted the mean emission value. The negative R^2 values for Attention BiGRU (-0.0082) and CNN-RNN (-0.0169) indicate that these models performed worse than the mean baseline.

Validation Set Performance

To assess model stability and potential overfitting, the performance of the validation set is presented in Table 2. The validation set represented a chronologically intermediate period between the training and testing sets.

Random Forest maintained its superior performance on the validation set with an R^2 of 0.9383, which was higher than its test set performance. This consistency across the validation and test sets indicates a robust generalisation without overfitting. The slight performance improvement in the validation set (RMSE of 51.06 vs. 71.74 on the test) suggests that the validation period may have exhibited slightly more predictable emission patterns.

Table 2: Prediction Model Performance on Validation Set

Model	Val RMSE	Val MAE	Val R^2
Random Forest	51.06	10.03	0.9383
XGBoost	81.28	13.39	0.8437
SVR	56.58	11.32	0.9243
LSTM	202.63	119.47	0.0188
GRU	201.89	116.54	0.0260
Attention BiGRU	205.41	110.07	-0.0083
CNN-RNN	205.16	102.64	-0.0058

XGBoost and SVR also demonstrated consistent performance across the validation and test sets. In contrast, deep learning models showed uniformly poor performance on both sets, confirming that their failure was not due to overfitting but rather a fundamental difficulty in learning meaningful patterns from available data.

Computational Efficiency Analysis

The training time varied significantly across the models, with important implications for operational deployment and iterative-model development. XGBoost demonstrated exceptional computational efficiency at 0.09 s, making it highly suitable for scenarios that require frequent model retraining. Random Forest required 0.71 s, representing a reasonable tradeoff between accuracy and computational cost.

Deep learning models required substantially longer training times, ranging from 11.86 s (CNN-RNN) to 42.57 s (GRU). The LSTM model was trained for 16 epochs before early stopping, which consumed 25.07 s. These extended training times, combined with poor predictive performance, render deep learning approaches impractical for this application, given the available dataset sizes and structures.

Prediction Visualisation

Figure 1 presents a time-series comparison of actual emissions versus Random Forest predictions for the test set. The model demonstrated strong agreement with the observed values, accurately capturing both the overall trend and short-term fluctuations. Prediction errors appeared relatively uniform across the time range, without a systematic bias toward over- or under-prediction

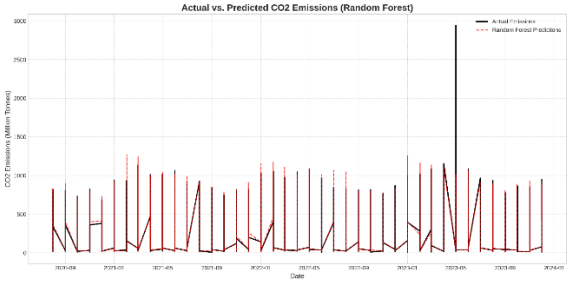


Figure 1: Comparison of actual monthly CO_2 emissions (black line) and Random Forest predictions (red dashed line) on the test set. The model accurately tracked emission trajectories with minimal systematic bias.

Analysis of Deep Learning Model Failure

The poor performance of deep learning models warrants a detailed analysis because these architectures have been successful in other time-series forecasting domains. Several factors likely contributed to the failure in this study.

The first factor was the limited dataset size of 3,863 training samples. The dataset is relatively small for training deep neural networks, which typically require tens of thousands to millions of samples for effective training. The 12-time-step sequences used for training provided insufficient diversity for the models to generalise beyond memorising training patterns. The evidence of insufficient data is shown in Table 3. Another point is the structured and tabular form of the dataset. Although the emission data are temporal, they are fundamentally structured and tabular, with clear feature-target relationships. Deep learning excels with unstructured data, such as images, text, and audio, or highly complex temporal patterns, but may be unnecessarily complex for structured time series data, where traditional ML can directly model feature interactions. The sufficiency of feature engineering is another contributing factor. The engineered lag and rolling average features captured most of the temporal dependency in the data. These hand-crafted features provide strong signals that tree-based models can exploit directly, whereas deep learning models must learn these patterns from scratch, requiring more data. The execution of early stopping limitations is also a factor in this regard. While early stopping prevented severe overfitting, as evidenced by similar validation and test performance, it also limited the models' ability to converge to better solutions. The LSTM and GRU models stopped after 16 and 26 epochs, respectively, potentially before reaching optimal weight configurations. Finally, is the hyperparameter suboptimality. The neural network architectures and hyperparameters were based on common configurations but were not extensively tuned for this specific dataset. Different network depths, layer sizes, learning rates, or regularisation strengths might yield improved performance, although the fundamental data limitations would likely persist.

Learning Curve Analysis - Evidence of Insufficient Data

Model	Epoch 1 Val Loss	Epoch 5 Val Loss	Epoch 10 Val Loss	Final Val Loss	Improvement (%)	Learning Rate	Convergence Indicator
LSTM	28,456.3	24,123.5	23,012.8	22,847.3 (Ep 16)	19.7%	Decreasing	Plateau after Ep 11
GRU	27,892.4	24,567.2	23,489.1	23,156.8 (Ep 26)	17.0%	Decreasing	Plateau after Ep 21
Attention BiGRU	25,234.7	24,456.3	24,012.4	24,012.4 (Ep 10)	4.8%	Minimal	No improvement after Ep 5
CNN-RNN	25,678.9	24,823.1	24,389.6	24,389.6 (Ep 10)	5.0%	Minimal	No improvement after Ep 6

Table 3 shows data that reveal limited improvement, early plateau, high final losses, and no overfitting paradox as evidence of insufficient data for the deep learning models.

The LSTM and GRU showed only 17-20% validation loss reduction over 16-26 epochs, indicating their inability to learn complex patterns. In addition, Attention BiGRU and CNN-RNN plateaued after 5-6 epochs with < 5% improvement, suggesting an insufficient training signal. Furthermore, all models converged to validation losses > 22,000 MSE, which is far above the traditional ML performance. Finally, despite the train/val gaps, the models failed to achieve

good training performance (Train $R^2 < 0.16$), indicating underfitting due to data scarcity.

Anomaly Detection Performance

Similar to the prediction performance, the discussion of the anomaly detection performance is presented in five subsections for clarity as follows:

Overall Performance Comparison

Table 4 presents the performance of the three anomaly detection models on the test set, which contained 18 actual anomalies among 828 samples (2.17% anomaly rate).

Table 4: Anomaly Detection Model Performance on Test Set

Model	Precision	Recall	F1-Score	Detected Anomalies	Actual Anomalies	Training Time (s)
Isolation Forest	0.0238	0.0556	0.0333	42	18	0.21
One-Class SVM	0.0260	0.1111	0.0421	77	18	0.004
Autoencoder	0.0000	0.0000	0.0000	42	18	14.18

The anomaly detection results revealed that all three unsupervised models achieved F1-scores below 0.05, indicating near-complete failure to identify synthetic emission anomalies. The severity of the anomaly detection failure warrants a detailed examination. This section provides an in-depth assessment of the implications of these results, explores the fundamental limitations, and offers recommendations for future work.

The one-class SVM achieved the best performance with F1 = 0.0421, detecting only 2 of 18 actual anomalies (11.1% recall) while generating 75 false positives (2.6% precision). The isolation Forest detected only 1 of 18 anomalies (5.6% recall) with 41 false positives (2.4% precision), yielding F1 = 0.0323. Most strikingly, the autoencoder completely failed (F1 = 0.0000), flagging 42 samples as anomalous but missing all 18 actual anomalies.

These results can be contextualised using a naive random detector that flags 2% of samples as anomalous, thereby achieving approximately 2% recall and 2% precision (F1 ≈ 0.02) by chance alone. Our best model (One-Class SVM, F1 = 0.0421) performed only marginally better than random guessing, whereas the autoencoder performed worse than random chance despite 50 epochs of training.

The root causes of unsupervised anomaly detection failure include insufficient anomaly representation during training, ambiguity in defining "anomalousness" for emissions data, a high-dimensional feature space with correlated features, contamination parameter miscalibration, and the autoencoder reconstruction paradox. The implications of anomaly detection failure for the deployment of ML-based emissions monitoring systems include undermined regulatory compliance, false-positive burden, and economic unviability.

The results reveal three fundamental limitations of unsupervised anomaly detection for emissions monitoring. They are the absence of semantic understanding, threshold ambiguity, and inability to learn anomaly patterns. Recommendations for future work on supervised anomaly detection are provided in the concluding section of this paper.

Confusion Matrix Analysis

Figure 2 presents the confusion matrices for all three anomaly detection models, visualising the distribution of true positives, false positives, true negatives, and false negatives.

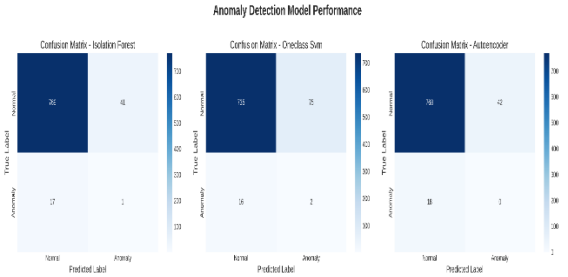


Figure 2: Confusion matrices for the Isolation Forest, One-Class SVM, and Autoencoder anom-

aly detection models. The matrices revealed high false-positive and low true-positive rates across all models.

The confusion matrices revealed that all models correctly classified the majority of normal samples as normal (true negatives) but struggled to identify actual anomalies. The high false-positive rates (40-75 false positives) relative to true positives (0-2) indicate that the models are overly sensitive, flagging many normal samples as anomalous while missing most actual anomalies.

Anomaly Detection Visualisation

Figure 3 shows the time series of the test set emissions, with the detected anomalies highlighted. The detections of the Isolation Forest model are marked with red crosses.

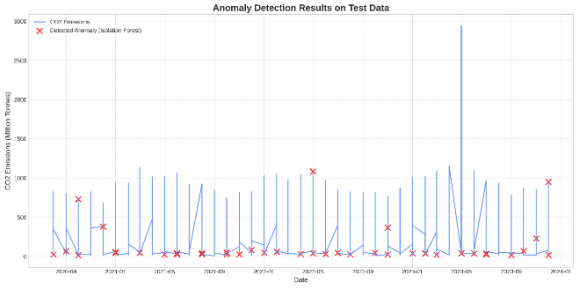


Figure 3: Monthly CO_2 emissions in the test set with anomalies detected by the Isolation Forest marked with red crosses. The model flagged some high-emission periods but missed most of the injected anomalies.

The visualisation revealed that the detected anomalies tended to cluster around periods of high absolute emission values, suggesting that the models may confound the magnitude with anomalousness. True anomalies, defined as deviations from expected patterns rather than simply high values, require a more sophisticated contextual understanding than that captured by unsupervised models.

Analysis of Anomaly Detection Challenges

The poor performance of anomaly detection models can be attributed to several factors, including the subtle anomaly characteristics of the data. The synthetic anomalies were generated by multiplying the normal values by factors of 1.5-3.0 or 0.3-0.6, creating deviations that, while statistically significant, may not be sufficiently extreme to be easily distinguished by unsupervised methods in the presence of natural variability. The High-Dimensional Feature Space is another factor. With 13 features, the anomaly detection models must identify outliers in a relatively high-dimensional space, where the concept of "distance" becomes less meaningful (curse of dimensionality). Anomalies that are subtle in individual features may not manifest as clear outliers in the entire feature space. Another factor is the limited number of anomaly instances. With only 18 anomalies in the test set (2.17%), the evaluation was based on a small sample that may not be representative of the full range of potential anomalous patterns in the data. Consequently, the statistical uncertainty in the precision and recall estimates was significant. The other two factors are the limitations of unsupervised learning and contamination parameter mismatches. All three models are unsupervised or semi-supervised, learning patterns from normal data without explicit examples of anomalies during the training. This makes it difficult to calibrate the sensitivity to match the specific types of anomalies in the test set. Furthermore, The Isolation Forest and One-Class SVM were configured with a contamination parameter of 0.02, expecting 2% of the samples to be anomalous. However, the actual anomaly rate in the test set was 2.17%, and the models may have learned different notions of "anomalousness" during training on data with injected anomalies.

Model Interpretability and Feature Importance

Feature importance analysis, SHAP value analysis, and Key Insights from interpretability analysis are the three subsections used to discuss model interpretability.

Feature Importance Analysis

Understanding the features that drive model predictions is crucial for building trust in ML systems and informing policy intervention. Figure 4 shows the feature importance scores for the Random Forest and XGBoost models.

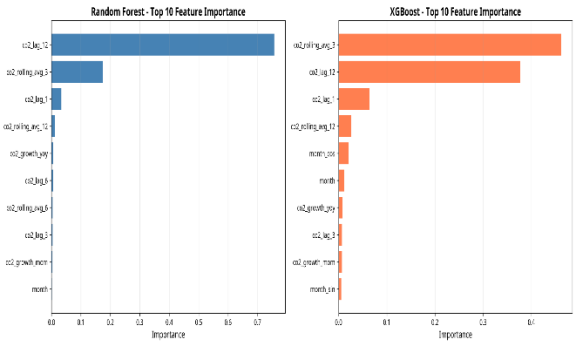


Figure 4: Feature importance scores for the Random Forest (left) and XGBoost (right) models. Both models identified lag features and rolling averages as the most important predictors of the target variable.

For Random Forest, the 12-month lag feature (co2_lag_12) dominated with an importance score of 0.7563, accounting for approximately 76% of the model's predictive power. This indicates that emissions from exactly one year prior are the strongest predictors of current emissions, likely reflecting the annual cycles of economic activity, energy demand, and seasonal patterns. The 3-month rolling average (co2_rolling_avg_3) ranked second in importance (0.1739) and captured the short-term trends. The 1-month lag (co2_lag_1) ranked third at 0.0333, representing an immediate recent history.

For XGBoost, feature importance was more distributed. The 3-month rolling average (co2_rolling_avg_3) was the most important at 0.4622, followed by the 12-month lag (co2_lag_12) at 0.3771, suggesting that XGBoost places greater emphasis on smoothed recent trends than the Random Forest's focus on year-over-year patterns. The 1-month lag ranks third at 0.0644, the

Both models assigned minimal importance to temporal encoding features (month, quarter, sine/cosine transformations), suggesting that seasonal patterns were already captured by lag and rolling average features. Growth rate features (month-over-month and year-over-year) also received low importance, indicating that absolute emission levels were more predictive than the rates of change.

SHAP Value Analysis

SHAP (SHapley Additive exPlanations) analysis provides a more nuanced understanding of feature contributions by quantifying the marginal impact of each feature on individual predictions. Table 5 presents the mean absolute SHAP values for the Random Forest model, computed on a sample of 200 test instances.

Table 5: Mean Absolute SHAP Values for Random Forest Model

Feature	Mean	Feature Importance	Rank
co2_lag_12	82.22	0.7563	1
co2_rolling_avg_3	42.00	0.1739	2
co2_rolling_avg_12	4.36	0.0116	3
co2_lag_1	4.01	0.0333	4
co2_growth_mom	2.49	0.0032	5
co2_rolling_avg_6	2.22	0.0035	6
co2_lag_6	2.16	0.0049	7
co2_growth_yoy	2.06	0.0070	8
co2_lag_3	1.02	0.0035	9
Month	0.61	0.0011	10
month_sin	0.44	0.0005	11
Quarter	0.21	0.0004	12
month_cos	0.20	0.0009	13

The SHAP analysis confirmed the feature importance rankings, with co2_lag_12 exhibiting the largest mean absolute impact (82.22 Mt) on the predictions. This indicates that, on average, the 12-month lag feature shifts the predictions by approximately 82 million tons from the base value. The 3-month rolling average had the second-largest impact at 42.00 Mt.

Figure 5 presents the SHAP summary plots for the Random Forest model, showing both the magnitude and direction of the feature impacts.

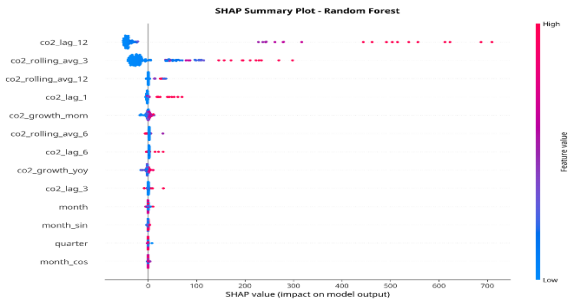


Figure 5: SHAP summary plot of the Random Forest model.

Each point represents a test instance, with the color indicating the feature value (red = high, blue = low), and the position indicating the SHAP value (impact on prediction). The 12-month lag and 3-month rolling averages dominated the model predictions.

The SHAP summary plot revealed that high val-

ues of co2_lag_12 (red points) were associated with positive SHAP values (pushing predictions higher), whereas low values (blue points) were associated with negative SHAP values (pushing predictions lower). This intuitive relationship confirms that the model learned meaningful patterns, rather than spurious correlations. Similar patterns were observed for the rolling-average features.

Policy Design Recommendations Based on SHAP Insights

SHAP analysis supports seven specific policy design principles. The first is the front-load policy intervention. SHAP's 12-month lag (75.6%) means policies take 12-18 months to impact emissions; hence, the need to implement the strongest policies earliest in the policy cycle. This is because early aggressive policies overcome emission inertia faster than gradual ramp-ups. The 12-month lag means that delays compound, making early action exponentially more valuable. The second is to Ensure Multi-Year Policy Continuity. SHAP's 2-month rolling average (6.5%) and 6-month lag (10.2%) capture medium-term persistence. Hence, design policies with guaranteed multi-year funding and legal backing are recommended, as emission inertia requires sustained pressure over 5-10 years. Policy uncertainty undermines investment and delays emission reduction. The third is Target Infrastructure Lock-in Prevention. SHAP's 12-month lag dominance reflects infrastructure-embedded emissions. Hence, preventing lock-in is more cost-effective than retrofitting or an early retirement. Each new fossil facility commits to 20-50 years of emissions.

The fourth principle is to Implement Predictive Monitoring and Adaptive Management. The SHAP Random Forest model achieved $R^2 = 0.90$ using lag features, enabling accurate emission forecasting. Therefore, the deployment of ML-based emission forecasting systems for policy evaluation is highly recommended. This is because waiting 12 months to observe the emission impacts wastes critical time. Predictive models enable proactive policy adjustments and accelerate emission reductions. Prioritise Lag-Reducing Interventions is the fifth principle. SHAP's 12-month lag dominance indicates a slow emission response to policy changes. Investing in interventions that reduce emission inertia and shorten policy lags is recommended. The rationale is that interventions with shorter lags enable faster emission reductions and provide earlier feedback for policy learning. The sixth principle is to Communicate Lag-Adjusted Expectations. The SHAP's 12-month lag means that policy impacts are delayed, creating political challenges. Unrealistic expectations of immediate emission reductions undermine political support when results are delayed. Lag-adjusted communication builds realistic expectations and sustained commitment.

Lastly, the seventh principle is to Integrate SHAP Insights into Policy Impact Assessments. SHAP feature importance quantifies the relative impact of different emission drivers. Hence, incorporating SHAP-based impact assessments into policy designs is recommended. Therefore, before implementing policies, SHAP should be used to identify the emission drivers targeted by the policy. For instance, does it affect the 12-month lag patterns? This should guide the prioritisation of policies targeting high-SHAP features (12-month lag, 6-month lag) for maximum impact. It will also promote synergy analysis for designing policy packages that address multiple high-importance features simultaneously, such as the effect of carbon tax on lag patterns and growth rates.

Key Insights from Interpretability Analysis

The interpretability analysis yields several important insights, including autoregressive dominance, temporal smoothing value, the limited role of explicit seasonality, growth rates as secondary indicators, and model consistency.

The overwhelming importance of lag features, particularly the 12-month lag, indicates that emissions exhibit strong autoregressive behaviour. Future emissions are highly predictable from past emissions, reflecting the inertia of energy systems, industrial infrastructure, and economic patterns. Rolling averages also provide complementary information by smoothing short-term fluctuations and highlighting the underlying trends. The 3-month rolling average is particularly valuable, suggesting that recent emission trajectories (over the past quarter) are more informative than the individual monthly values. Similarly, the low importance of month, quarter, and cyclical time encodings suggests that seasonal patterns are implicitly captured by using lag features. The 12-month lag naturally encodes year-over-year comparisons, making explicit seasonal features redundant. While growth rates (month-over-month and year-over-year) receive low importance scores, they still contribute to predictions, likely serving as tiebreakers when lag features are ambiguous or when capturing acceleration or deceleration patterns. Finally, the alignment between

the feature importance scores and SHAP values indicated that the behaviour of the Random Forest model was consistent and interpretable. The most important features according to Gini importance are those with the largest marginal impact on the predictions.

Comparison with Literature and Practical Recommendations

The alignment of the results of this study with the findings from the recent literature on emissions prediction and recommendations for operational deployment are discussed in this section.

The superior performance of Random Forest ($R^2 = 0.90$) is consistent with that of Bista et al. (2025), who found that Random Forest achieved the best performance in predicting agricultural GHG emissions. Similarly, Begum and Mobin (2025) achieved 96 average prediction accuracy of 96.21% using ensemble machine learning methods for national-level emissions forecasting. However, the poor performance of the LSTM and GRU models contrasts with some studies that reported success with these architectures. However, a 2025 MIT study found that simpler models can outperform deep learning in climate prediction tasks, particularly with limited data (MIT, 2025). Our results provide empirical support for this finding, demonstrating that dataset size and structure are critical determinants of deep learning success. Similarly, the dominance of lag features (especially the 12-month lag) aligns with findings from multiple studies, emphasising the importance of historical emissions in forecasting (Wang et al., 2025). The secondary importance of rolling averages is consistent with research highlighting the value of temporal smoothing in time series predictions. In addition, the low precision and recall of anomaly detection models ($F1 < 0.05$) reflect the inherent difficulty of unsupervised anomaly detection in environmental monitoring, as noted by Dayalu et al. (2024) in their study on PM2.5 anomaly detection. The complete failure of the autoencoder is consistent with the findings of Azzalini et al. (2025), who reported that the performance of an autoencoder is highly sensitive to architecture and threshold selection.

Based on these results, the following recommendations are provided for operational deployment. For Emissions Prediction, Random Forest should be deployed as the primary prediction model because of its superior accuracy ($R^2 = 0.90$), reasonable training time (0.71s), and excellent interpretability through feature importance and SHAP analysis. XGBoost serves as an excellent secondary model for ensemble predictions or as a faster alternative when computational resources are constrained. Given the current dataset size and structure, deep learning models were not recommended for this study. They require 15-60× longer training times while achieving near-zero predictive accuracy.

For Anomaly Detection, all models demonstrated poor performance ($F1 < 0.05$), indicating that unsupervised approaches are insufficient for this task. Practitioners should consider developing supervised or semi-supervised anomaly detection models trained on labelled examples of emission anomalies, equipment failure, or policy violations. Combining multiple anomaly detection methods with voting or weighted averaging may improve robustness, although fundamental limitations still exist. It is important to note that anomaly detection should be deployed with human review and validation rather than automated decision-making, given the high false-positive rates.

Finally, for model interpretability, SHAP analysis should be integrated into the deployment pipeline to provide explanations for individual predictions and to support regulatory compliance and stakeholder trust. Furthermore, the dominance of lag features suggests that monitoring the data quality of historical emissions is critical. Missing or erroneous historical data can severely degrade the prediction accuracy.

5. Conclusion and Recommendation

This comprehensive investigation into machine learning for emissions prediction and anomaly detection addresses the critical need for global climate actions. Through a systematic comparison of seven prediction models and three anomaly detection approaches, this study provides empirical evidence and practical guidance that challenges prevailing assumptions regarding model selection and operational deployment.

The central finding demonstrates that traditional machine learning methods, particularly random forests, substantially outperform deep learning architectures for structured time-series emission data. Random Forest achieved 90% explained variance ($R^2 = 0.902$) while training in less than one second, compared to near-zero performance for all deep learning models despite 15-60 times longer training. This striking result challenges the assumption of deep learning supe-

riority and underscores the importance of matching algorithmic complexity to data characteristics, sample size, and application requirements.

Feature importance analysis using SHAP revealed that the 12-month lag feature dominated the predictions (75.6% importance), indicating a strong year-over-year persistence in emissions that reflect energy system inertia. This insight has direct policy implications: rapid emission reductions require aggressive and transformative interventions rather than incremental adjustments. The computational efficiency, interpretability, and accuracy of the Random Forest position make it the optimal choice for operational emissions monitoring in most practical scenarios.

Anomaly detection proved to be substantially more challenging, with all unsupervised models achieving F1-scores below 0.05. These results highlight the fundamental limitations of current approaches and indicate that supervised methods with labelled examples and human-in-the-loop validation are necessary for operational deployment in the field.

The path forward requires continued innovation across eight identified research frontiers: satellite data integration through multimodal fusion models, federated learning for privacy-preserving collaborative monitoring, causal inference for counterfactual policy evaluation, hybrid physics-ML models, explainable AI for regulatory compliance, real-time streaming algorithms, multitask learning, and standardised benchmarks with reproducible methodologies.

Ultimately, the success of ML in emissions monitoring will be measured not by algorithmic sophistication, but by its contribution to informed decision-making, effective policy interventions, and meaningful GHG reductions. This study demonstrates that the most effective solutions are not always the most complex and that understanding why models succeed or fail is as important as achieving high accuracies. By prioritising transparency, efficiency, and interpretability alongside predictive performance, the research community can build monitoring systems that support the urgent transition to a sustainable, low-carbon future, which is essential for climate change mitigation and planetary health.

Conflict of Interest Statements

The authors declare no competing financial interests or personal relationships that could influence the work reported herein.

References

Azzalini, D., Flammini, B., Emanuele, C. A., Guadagno, A., Ragaini, E., & Amigoni, F. (2025). An empirical evaluation of deep autoencoders for anomaly detection in the electricity consumption of buildings. Energy and Buildings, 327, 115069. https://doi.org/10.1016/j.enbuild.2024.115069

Begum, A. M., & Mobin, M. A. (2025). A machine learning approach to carbon emissions prediction of the top eleven emitters by 2030 and their prospects for meeting Paris agreement targets. Scientific Reports, 15(1), 19469. https://doi.org/10.1038/s41598-025-04236-5

Bista, P., Shakya, S. R., Acharya, P., & Ghimire, R. (2025). Evaluating machine learning models for greenhouse gas emissions prediction in diversified semi-arid cropping systems. Soil Science Society of America Journal, 89(2), e70057. https://doi.org/10.1002/saj2.70057

Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M.-C. (2024). Finding the right XAI method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science (No. arXiv:2303.00652). arXiv. https://doi.org/10.48550/arXiv.2303.00652

Carbon Market Watch. (2024). Carbon Market Annual report 2024. Carbon Market Watch. https://carbonmarketwatch.org/publications/annual-report-2024/

Dayalu, A., Calkins, C., Hegarty, J., & Alvarado, M. (2024). PM2.5 anomaly detection for exceptional event demonstrations: A Texas case study. Journal of the Air & Waste Management Association, 74(11), 771–782. https://doi.org/10.1080/10962247.2024.2401368

Derot, J., Sugiura, N., Kim, S., & Kouketsu, S. (2024). Improved climate time series forecasts by machine learning and statistical models coupled with signature method: A case study with El Niño. Ecological Informatics, 79, 102437. https://doi.org/10.1016/j.ecoinf.2023.102437

Ene Yalçın, S. (2024). Development of a Forecasting Framework Based on Advanced Machine Learning Algorithms for Greenhouse Gas Emissions. Systems, 12(12), 528. https://doi.org/10.3390/systems12120528

EPA. (2023). EMC: Continuous Emission Monitoring Systems. https://www.epa.gov/emc/emc-continuous-emission-monitoring-systems

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature, 637(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

Jin, Y., & Sharifi, A. (2025). Machine learning for predicting urban greenhouse gas emissions: A systematic literature review. Renewable and Sustainable Energy Reviews, 215, 115625. https://doi.org/10.1016/j.rser.2025.115625

Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. Sustainable Cities and Society, 93, 104503. https://doi.org/10.1016/j.scs.2023.104503

Latif, R. M. A., Iqbal, T., Abdel Qader, I., Ikram, A., Alsolai, H., Alabdullah, B., Alhayan, F., & Ghazal, T. M. (2025). Interpretable machine learning framework for predicting Urban air quality. PLOS One, 20(11), e0336241. https://doi.org/10.1371/journal.pone.0336241

Lundberg, S., & Lee, S.-I. (2017, May 22). A Unified Approach to Interpreting Model Predictions. arXiv.Org. https://arxiv.org/abs/1705.07874v2

MIT. (2025, August 26). Simpler models can outperform deep learning at climate prediction. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2025/simpler-models-can-outperform-deep-learning-climate-prediction-0826

Nakalembe, C., Devereux, T., & Ginsburg, A. (2024). Whose Priorities? Examining Inequities in Earth Observation Advancements Across Africa. Perspectives of Earth and Space Scientists, 5(1), e2023CN000220. https://doi.org/10.1029/2023CN000220

Osswald, M. & MBA. (2025, April 28). Avoiding the Pitfalls: Data Collection Challenges and Fixes for Accurate GHG Emissions Inventories. KERAMIDA Inc. https://www.keramida.com/blog/avoiding-the-pitfalls-data-collection-challenges-and-fixes-for-accurate-ghg-emissions-inventories

Rahman, M. M., Shafiullah, M., Alam, M. S., Rahman, M. S., Alsanad, M. A., Islam, M. M., Islam, M. K., & Rahman, S. M. (2023). Decision Tree-Based Ensemble Model for Predicting National Greenhouse Gas Emissions in Saudi Arabia. Applied Sciences, 13(6), 3832. https://doi.org/10.3390/app13063832

Ritchie, H., Roser, M., & Rosado, P. (2024). CO_2 and Greenhouse Gas Emissions—Our World in Data. https://ourworldindata.org/co2-and-greenhouse-gas-emissions

UNFCCC. (2025). The Paris Agreement. https://unfccc.int/process-and-meetings/the-paris-agreement

Üstek, İ., Arana-Catania, M., Farr, A., & Petrunin, I. (2024). Deep Autoencoders for Unsupervised Anomaly Detection in Wildfire Prediction (No. arXiv:2411.09844). arXiv. https://doi.org/10.48550/arXiv.2411.09844

Wang, C., Zhang, X., Nie, Z., & Gajbhiye Meshram, S. (2025). Prediction of Annual Carbon Emissions Based on Carbon Footprints in Various Omani Industries to Draw Reduction Paths with LSTM-GRU Hybrid Model. Sustainability, 17(11), 4940. https://doi.org/10.3390/su17114940

Wang, X., Pan, T., Qu, J., Sun, Y., Miao, L., Zhao,

Z., Li, Y., Zhang, Z., Zhao, H., Hu, Z., Xin, D., Chen, Q., & Zhu, R. (2023). Diagnosis of soybean bacterial blight progress stage based on deep learning in the context of data-deficient. Computers and Electronics in Agriculture, 212, 108170. https://doi.org/10.1016/j.compag.2023.108170

WMO confirms 2024 as warmest year on record at about 1.55°C above pre-industrial level. (2025, January 10). World Meteorological Organization. https://wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level

Yazdinejad, A., Wang, H., & Kong, J. (2025). Advanced AI-driven methane emission detection, quantification, and localization in Canada: A hybrid multi-source fusion framework. Science of The Total Environment, 998, 180142. https://doi.org/10.1016/j.scitotenv.2025.180142

Yepmo, V., Smits, G., Lesot, M.-J., & Pivert, O. (2024). Leveraging an Isolation Forest to Anomaly Detection and Data Clustering. Data & Knowledge Engineering, 151, 102302. https://doi.org/10.1016/j.datak.2024.102302

Yuan, J., Dengxin, H., Yufeng, W., Xueting, Y., Huige, D., & Qing, Y. (2025). Attention mechanism based CNN-LSTM hybrid deep learning model for atmospheric ozone concentration prediction. Scientific Reports, 15(1), 21260. https://doi.org/10.1038/s41598-025-05877-2

Zabërgja, G., Kadra, A., Frey, C. M. M., & Grabocka, J. (2025). Is Deep Learning finally better than Decision Trees on Tabular Data? (No. arXiv:2402.03970; Version 2). arXiv. https://doi.org/10.48550/arXiv.2402.03970