RESEARCH ARTICLE

A Data-Driven Framework for Methane Emission Prediction Using Machine Learning Methods

Suraj Arya[1], **Anju**[1,*] **and Jonas Nnaemeka Onah**[2]

[1] Department of Computer Science and Information Technology, Central University of Haryana, Haryana 123031, India
[2] Department of Electrical and Electronics Engineering, Federal University of Petroleum Resources, Effurun, Delta State 320102, Nigeria

Abstract

Greenhouse gas Methane (CH_4) has 86 times more impact on global warming than carbon dioxide (CO_2). The emission of methane gas into the atmosphere is increasing due to the reliance on fossil-based resources in post-industrial energy consumption, along with the rise in food demand and the generation of organic waste that accompanies a growing human population. CH_4 acts as a vital pollutant in the air. The problem addressed in this study was to accurately estimate CH_4 emissions from functional urban areas. This study aims to predict CH_4 emissions using Time Series (TS) and Machine Learning (ML) models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Long Short-Term Memory (LSTM), Random Forest Regressor (RFR), and CatBoost Regressor (CABR), etc. The SARIMA model has the best combination of values (1,0,0) (1,1,0). The methane emission data was collected from the World Bank's Group from 2019 to 2022. Among all models, the SARIMA model predicted CH_4 emissions more accurately than the other models. The results obtained in the study indicate that SARIMA outperforms other techniques. The SARIMA model performed the most accurate results in terms of R-squared score (R^2) = 94%; Root Mean Squared Error (RMSE) = 2.8129; Mean Squared Error (MSE) = 7.9126; Mean Absolute Error (MAE) = 1.8391, etc. This type of prediction enables the government to reduce CH_4 emissions at the global level.

Keywords: machine learning, methane emission, random forest regressor, prediction.

1 Introduction

Although the presence of methane in the atmosphere is short-lived, it is a highly potent greenhouse gas (GHG), with a global warming potential 28 times higher than carbon dioxide over a 100-year period. Its emissions have far-reaching consequences, impacting the environment, human health, and the economy [1, 2]. Prolonged exposure to methane and other air pollutants increases the risk of certain cancers. Economically, methane emissions can lead to significant losses in productivity and hinder economic growth due to their detrimental effects on human health and the environment [3]. Furthermore, research has shown that methane emissions can compromise infrastructure integrity, damage pipelines and buildings, and resulting in costly

Submitted: 22 October 2025
Accepted: 25 December 2025
Published: 12 January 2026

Vol. 2, **No.** 1, 2026.
10.62762/TMI.2025.782852

***Corresponding author:**
✉ Anju
anju24sanga@gmail.com

Citation
Arya, S., Anju, & Onah, J. N. (2026). A Data-Driven Framework for Methane Emission Prediction Using Machine Learning Methods. *ICCK Transactions on Machine Intelligence*, 2(1), 53–64.

repairs and maintenance [4]. Nigeria on one hand, ranked first in Africa and among the top 10 countries for gas flaring globally in 2020, with approximately 7 billion cubic meters of gas wasted [5, 6]. However, according to the World Bank's Global Gas Flaring Tracker Report, Nigeria has made significant progress in reducing gas flaring, achieving a 70% reduction to 7 billion cubic meters (bcm) in 2020 compared to previous years. Methane emission is significantly higher in the oil and gas producing countries. The majority of these emissions (73%) came from gas production, processing, and distribution, while 27% came from oil production in Nigeria. In 2010, Nigeria's oil and gas sector emitted an estimated 439.8 kilotonnes of methane. The projections indicate a significant increase in methane emissions at 481.2 kilotonnes by 2030 (9% increase) and 598.5 kilotons by 2050 (36% increase) [7]. On the other hand, China ranks first in Methane emission, and is responsible for nearly one-fifth of global methane emissions [8]. Historically, methane wasn't a major focus in China's climate policies until the early 2010s. However, in recent years, China has prioritized reducing anthropogenic methane emissions, incorporating it into both domestic policies and international commitments [9].

Various methods exist for measuring methane emissions, including ground-based, aircraft, and satellite-based monitoring [10]. Meanwhile, researchers are increasingly leveraging machine learning techniques to estimate methane emissions, offering new opportunities for accurate measurement and analysis [11]. The work by [12] leveraged the deep learning-based approach like convolutional neural network to quantify methane emission for field application. Although, the objectives set out by the authors were achieved. However, the approach is limited to unstructured data set (images and texts) in its application. The work by [4] leveraged a deep learning approach, specifically convolutional neural networks (CNNs), to estimate gas emissions. While their model showed promising performance, further refinement is needed to achieve more accurate results. In the work of [13] there is a clear and concise comparison of models, with promising results for air quality forecasting. The study compares three distinct models Long Short-Term Memory (LSTM) recurrent neural network, Fully-Connected Neural Network (FC-NN), and Autoregressive Integrated Moving Average (ARIMA), providing a comprehensive evaluation, indicating that LSTM outperforms the other two methods. However, the reasoning behind LSTM's superior performance could be explored further and hyperparameter tunning is a major concern for each of the cases. The study's findings can inform the development of cost-effective predictive emissions monitoring systems. While the study compares model performance, further analysis of the models' interpretability could enhance understanding. Again, the study might benefit from a more detailed discussion on hyperparameter tuning and its impact on model performance.

Literature [14] compare machine learning models to estimate wetland methane emissions, offering a robust and data-driven approach. The use of a multi-model ensemble approach helps reduce uncertainties and improves the reliability of estimates. The research takes into account a number of factors, such as climate type, wetland types, soil characteristics, air temperature, and precipitation. While the multi-model ensemble (MME) approach reduces uncertainties, further research could explore additional methods to minimize errors. More detailed validation of the models against independent datasets would strengthen the study's findings. In [15] the work combined machine learning, satellite data, climate data, and production data to predict ground-level methane concentrations. It compares the performance of different machine learning models, identifying the Extreme Gradient Boost model as the most accurate. While the Extreme Gradient Boost model performs well, its complexity might limit interpretability. A study by [16] employed machine learning (ML) models to estimate methane emissions from 97,435 Chinese reservoirs, categorized by storage capacity. The comprehensive assessment estimated total emissions at approximately 5,414 Gg, with reservoirs larger than 0.01 km^3 accounting for around 90% of emissions due to high diffusive flux rates and extensive surface areas. Thermal stratification and organic matter accumulation contributed to elevated methane diffusion in these reservoirs. However, the models' complexity may limit interpretability, and the findings might be specific to Chinese reservoirs, potentially limiting their applicability to other regions. The following are the primary contributions of this study:

- The aim of this study was to predict CH$_4$ emission of Functional Urban Area (FUA) using satellite data, machine learning regression and time series (TS) models. The study focuses on time series forecasting from 2019 to 2022.

Table 1. Research gap of related work.

S.No	Ref.	Dataset used/ Data source (Name)	Region Focus (City / Country)	Methods Used for Analysis (AI / ML etc.)	Research Results (e.g., accuracy)	Identified Research Gap	Future Scope
1	[17]	Methane and hydrogen blends	United Kingdom	Ensemble learning algorithm	3.453e+04 STD	More hyperparameters to tune introduces complexity in the selection process	It can be integrated with automated machine learning to automate the process of building and selecting ensemble models.
2	[18]		United Kingdom	Multi-Layer Perceptron	MSE occurs at λ = 0.001. RMSE values LV 5	May get stuck in local minima:	Ensemble methods
3	[19]	AER CH_4 methane concentration of real time data from air monitoring station	Alberta	LSTM- ANN	RMSE = 0.1268 MAPE = 2.4134%	Difficulty in interpretation of the relationships between variables.	Incorporate data from satellite imagery, sensor networks to improve model accuracy and generalizability.
4	[20]	Intergovernmental Panel on Climate Change with satellite-based measurements from Sentinel-5P	Countries listed in Annex I of United Nations	Trend analysis	R^2 value for both the testing (R^2 = 0.973), and the training (R^2 = 0.981)	complexity in fine-tuning the model to identify optimal process conditions for Hydrogen Sulphide Methane Reformation (HSMR)	Optimize the ANN surrogate model by exploiting evolutionary algorithms.
5	[21]	TROPOMI and GOSAT data	China	UNMAMO algorithm with random forest model	R^2 = 0.91 RMSE = 17.16	Intensive computing requirement	Simple computation requirement
6	[22]	Satellite data	Taiwan	UNMAMO	Accuracy = 97.2% and an R^2 score of 0.858	Enhanced accuracy and efficiency compared to the traditional method	Despite the fact that it has large computing burden, the inherent constraints due to the algorithm could limit effective methane monitoring solutions.
7	[23]	Real time data	Surat Basin in Australia,	Multi-channel dynamic LSTM-ANN		Enhanced accuracy in forecasting production levels by analyzing historical data patterns and trends in coalbed methane extraction.	
8	[24]	Surface rates and pressure data.	China	BO-LTSM		It relies on a particular input data with potential requirement of validation to generalize the findings.	
9	[25]	Real time data	China	Temporal Convolutional Network (TCN)	MSE, MAE, RMSE, and R^2	No explicit information to validate claim	Simpler, accurate and less time consuming approach is required

- Recent studies use ML only for datasets that are already available in the same form. But our study integrates the yearly dataset to make a more diverse dataset and used ML as well as time series models for CH_4 emission prediction. This makes our study novel. The SARIMA achieved an R^2 score 94
- The dataset is novel and it is not yet analyzed.

To systematically identify the limitations of current approaches and position our contributions, a comprehensive research gap analysis of related work is presented in Table 1. The table compares recent studies across key dimensions such as dataset, methodology, region, and identified limitations.

Accurate CH_4 emissions estimation remains a noteworthy challenge due to intricate and dynamic behavior of CH_4 sources in FUA. Some traditional methods fail to accurately estimate CH_4 emissions. Additionally, the dataset used in this study suffer from noise, null value, duplicate values, and outliers which leads to inaccurate estimates of CH_4 emissions. Therefore, to address this problem, we utilized ML and TS models to improve CH_4 emission estimation, ensuring more reliable predictions.

The current study focuses solely on CH_4 emissions from activities occurring in functioning urban areas. Our dataset spans the years 2019–2022 (limited time period), but to improve accuracy, future studies should include real-time observations and a more diverse dataset. For our dataset, SARIMA achieved the best results; hence, to estimate fluctuations in CH_4 emissions, future research should investigate ensemble learning and deep learning approaches.

2 Methodology

This study forecasts global methane emissions by utilizing various regression and time series methods of machine learning. To tackle methane emissions estimates for Functional Urban Areas (FUAs), the initial phase is business comprehension, involving the identification of the problem, understanding business viewpoints and requirements, and strategizing to achieve objectives. This involves offering details that link grid cells in the World Bank's global CH_4 database to identifiers for FUAs and national administrative entities.

Table 2. Dataset description.

Sr.No.	Feature Name	Description
1	id5	Grid cell unique ID (It may be pixel or location ID)
2	x	Represents the longitude of the observation point
3	y	Represents the Latitude of the observation point
4	terr5	The code urban indicator mostly values 1
5	mean_ch4	Mean methane concentration (parts per billion)
6	mean_ch4_anomaly	Deviation/Anomaly of average methane levels
7	year	Observation year (2019 to 2022)
8	month	Observation month (1 to 12)
9	fua_id	Functional Urban Area (FUA) identifier
10	max_coverage_fraction	Maximum coverage fraction for data
11	eFUA_name	Functional Urban Area (FUA) name
12	Cntry_ISO	Country code 3-letter, i.e., IND for India
13	Cntry_name	Country name (e.g., "India")
14	FUA_p_2021	Population of Functional Urban Area (FUA) in 2015

2.1 Selection of Machine Learning (ML) and Time Series (TS) models

ML is a computer technique that allows systems to recognize patterns in data and classify or predict outcomes of methane emission. ML is the scientific study of creating models, algorithms, and learning strategies that enable computers to learn in a manner like humans. The TS model ARIMA, SARIMA (Seasonal ARIMA), LSTM and different ML algorithms such as Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), AdaBoost Regressor (ADBR), CatBoost Regressor (CABR), LightGBM Regressor (LGBMR), and XGBoost Regressor (XGBR) are used to predict methane emission. The selection of these diverse models addresses the methodological limitations identified in Table 1.

2.2 Platform and dataset details

Anaconda Navigator was used as a platform for implementing machine learning and time series models in this study. The yearly dataset is downloaded from the World Bank Group (i.e., from 2019 to 2022). We used the CH_4 dataset from the European Space Agency's Sentinel-5P (S-5P) (TROPOMI) satellite platform [28]. The year and month columns indicate the temporal resolution of the dataset, as the data was collected monthly based on observations from 2019 to 2022. Columns "x", "y", and "id5" in Table 2 describe the spatial resolution of CH_4 dataset. Each yearly dataset contains fourteen features. After combining, the total rows become 1249411. Access to relevant data is necessary for successful data analysis. The dataset belongs to 179 countries with 4776 different FUA. The sample of the methane emission dataset is given in Table 2. The dataset is downloaded from the URL [1]

2.3 Dataset preprocessing

Data Cleaning: In this phase, missing, null, and duplicated values are removed from the dataset. We used interpolation to estimate missing data. After that, separate the numerical and categorical features of the methane emission dataset. In this dataset, there are three categorical columns and eleven numerical columns. Table 3 presents the statistics of categorical features of CH_4 dataset. From the table the unique value for Cntry_name is '179' and top is 'China'.

Table 3. Statistical description of categorical variable.

Index	eFUA_name	Cntry_ISO	Cntry_name
count	1249411	1249411	1249411
unique	4776	179	179
top	Wenshang	CHN	China
freq	15442	322394	322394

Data wrangling: In the data wrangling stage, convert the raw data, i.e., the year and month columns into structured data/usable format, i.e., the date column. In this way, fifteen features became. So, after removing both columns ("year" and "month"), thirteen features remain in the dataset.

Outlier: Outlier are the abnormal data point, which are different from normal data [29]. The outlier of all features is removed using the z-score method. Before and after outlier mean_ch4 feature shown in Figure 1.

Feature selection: After removing the outlier, descriptive statistics and important feature of

[1] https://datacatalog.worldbank.org/search/dataset/0064329/Methane-emissions-for-functional%20urban-areas.

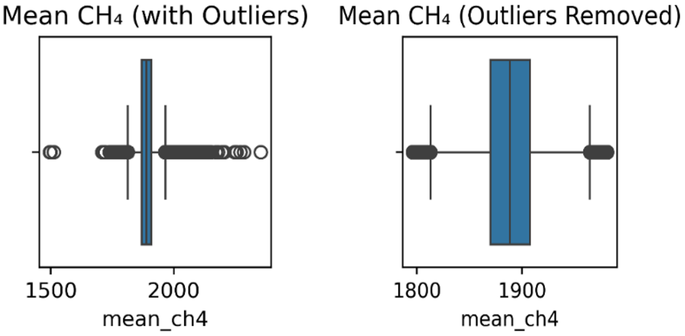

Figure 1. Target feature with and without outlier.

dataset was extracted for further analysis. Upon deep analysis of the dataset, we identified that only five relevant features will be used for predicting methane emissions. Others were excluded from the dataset. The independent variables “date”, "max_coverage_fraction", "FUA_p_2015", “mean_ch4_anomaly”, and the dependent variable ’mean_ch4’ were used in this study.

Dataset splitting: After feature selection, the dataset is split into training and testing using train_test_split (70:30, train:test) function.

Model selection: The next step is training a suitable model and evaluate the performance of models using different error metrics such as R-squared, Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Adjusted R^2 Score, Explained Variance Score (EVS), Mean Squared Log Error (MSLR), Median Absolute Error (MeAE), D2 Absolute Error Score (D2AES), D2 Pinball Score (D2PS), D2 Tweedie Score (D2TS), Mean Poisson Deviance (MPD), Mean Gamma Deviance (MGD), and Mean Tweedie Deviance(MTD). After this, the results were visualized. These steps of preprocessing are given in Figure 2.

RFR is an ensemble learning method that combines many decision trees to increase accuracy, as shown in Figure 3. It is applied to problems involving regression and classification. It selects features at random using the bagging, boosting, and stacking techniques of the ensemble. This algorithm was tuned using various parameters [26].

ADBR is an ensemble approach that uses a decision tree as the foundation model and the boosting technique. Because weights are given to each occurrence and greater weights are reassigned to instances that are mistakenly identified, it is referred to as adaptive boosting. The number of base models and the learning rate are parameters used in this model [27]. Compared to the other models, ADBR is less likely to overfit, and it may be used with other models to enhance performance, as shown in Figure 4.

LGBMR is an ML algorithm that is used for solving supervised learning regression tasks. It builds DT sequentially. The mathematical formula of LGBMR is given as:

$$B = F_N(A) = \sum_{n=1}^{N} \gamma_n \, h_n(A) \tag{1}$$

where: B: Predicted Final Value, A: Input features, N: number of DT, γ_n: Applied learning rate or weight, $h_n(A)$: n^{th} decision tree prediction.

ARIMA and SARIMA are statistical model used for methane forecasting. ARIMA is represented as ARIMA (p, d, q), where ’p’ stands for autoregressive (AR), ‘d’ means integrated (I), and ’q’ refers to the moving average (MA) component whereas SARIMA is represented as SARIMA (p, d, q) (P, D, Q)s, where seasonal components are (P, D, Q) and ‘s’ is the seasonal cycle length . The algorithm steps to predict methane emission using ARIMA and SARIMA are given below:

1. Load the methane emission dataset
2. Cleaning the dataset
3. Check dataset stationarity using the ADF (Augmented Dicky Fuller) test.
   (a) If the series is stationary, proceed. (Not Stationary shown in Figure 5)
   (b) Otherwise, make the series stationary by first-order differencing. (After differencing the stationary series is shown in Figure 5)
4. Generate PACF (Partial autocorrelation function) and ACP (Autocorrelation function), and find respective values of p and q. The ACF and PACF plot is depicted in Figures 6 and 7, respectively.
5. Train the ARIMA and SARIMA models.
6. Predict the methane emission for next years.

LSTM is a recurrent neural network that creates enhanced neural connections when the input data are arranged in a sequence. Since HSD is sequential, LSTM is a perfect deep learning model that leverages the sequence input. The LSTM model can retain significant information over extended periods using

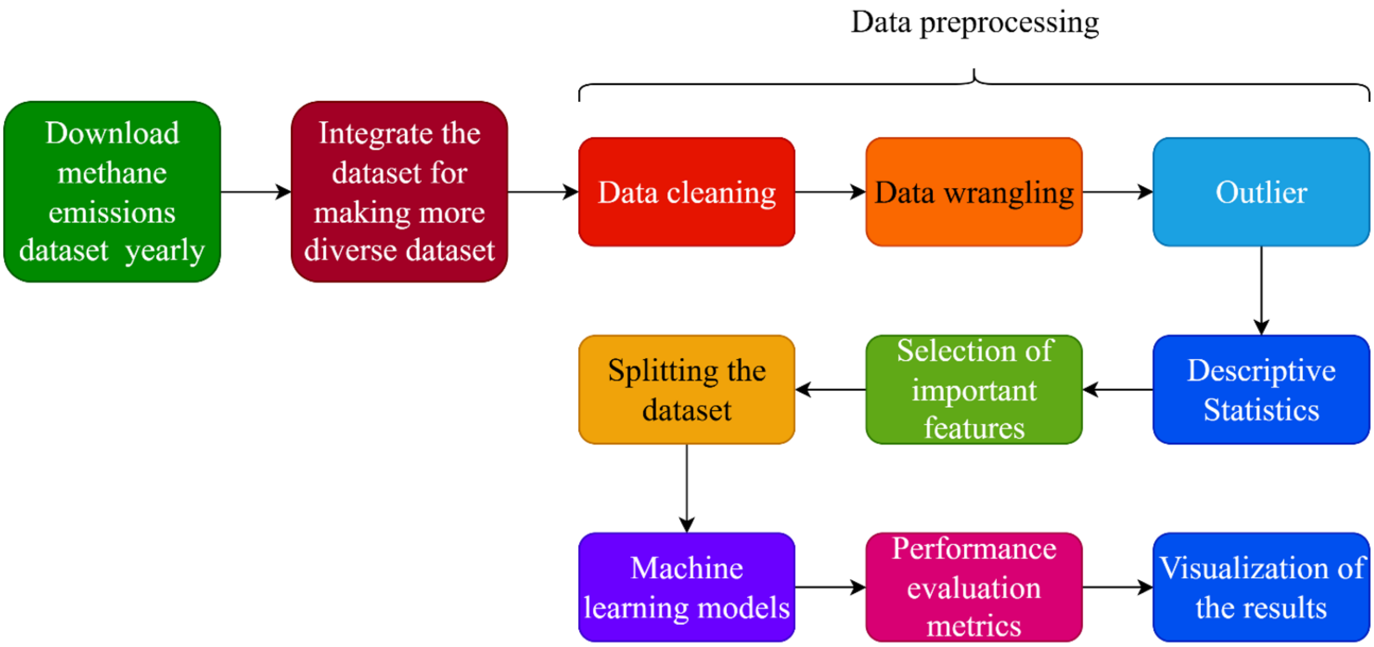
Data preprocessing
Download methane emissions dataset yearly
Integrate the dataset for making more diverse dataset
Data cleaning
Data wrangling
Outlier
Descriptive Statistics
Selection of important features
Splitting the dataset
Machine learning models
Performance evaluation metrics
Visualization of the results

Figure 2. Flow diagram of the implementation of ML models.

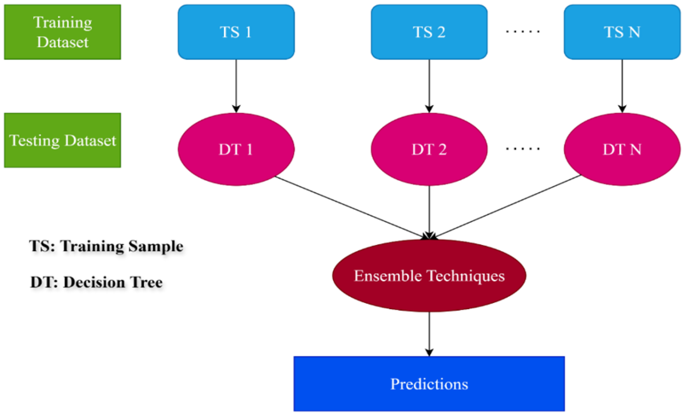

Figure 3. Random Forest.

gates. The LSTM model consists of four elements: (a) The Cell State serves as memory storage. (b) The forget gate eliminates irrelevant information. (c) In the cell state, the input gate decides what new information is added. The sigmoid function sigma helps control the information flow within gates. (d) The output gate decides which portion of the cell state will be utilized to produce the output, as illustrated in Figure 8.

3 Experimental design

3.1 Model performance evaluation metrics

The performance of ML and TS models were validated using the evaluation metrics shown in Table 4. These are R-squared, MSE, MAE, RMSE, MAPE, Adjusted R-squared Score, EVS, MSLR, D2AES, D2PS, D2TS, MPD, MGD, and MTD.

Table 4. Error metrics with their respective formula.

Error Name	Metrics	Formula used
MSE		$\frac{1}{m}\sum_{j=1}^{m}(X-Y)^2$
MAE		$\frac{1}{m}\sum_{j=1}^{m}\lvert X-Y\rvert$
RMSE		$\frac{1}{m}\sum_{j=1}^{m}(X-Y)^2$
MAPE		$\frac{100\%}{m}\sum_{j=1}^{m}\left\lvert\frac{X-Y}{X}\right\rvert$
Adjusted R^2 Score		$1-(1-R^2)\cdot\frac{m-1}{m-k-1}$
EVS		$1-\frac{\mathrm{var}(x-y)}{\mathrm{var}(x)}$
MSLE		$\frac{1}{m}\sum_{j=1}^{m}(\log(1+X)-\log(1+Y))^2$
D2AES		$1-\frac{\sum_{j=1}^{m}\lvert X-Y\rvert}{\sum_{j=1}^{m}\lvert X-Z\rvert}$
D2PS		$1-\frac{\sum L(X,Y)}{\sum L(X,Z)}$
D2TS		$1-\frac{\text{tweedie deviance}(x,y)}{\text{tweedie deviance}(x,z)}$
R^2		$1-\frac{\sum_{j=1}^{m}(X-Y)^2}{\sum_{j=1}^{m}(X-Z)^2}$
MPD		$\frac{1}{m}\sum_{j=1}^{m}\left[X\log\left(\frac{Y}{\bar{Y}}\right)-(X-Y)\right]$
MGD		$\frac{1}{m}\sum_{j=1}^{m}\left[X\log\left(\frac{Y}{\bar{Y}}\right)-\frac{X-Y}{Y}\right]$
MTD		$\frac{1}{m}\sum_{j=1}^{m}D(X,Y)$

where:

X = Actual value, Y = Predicted value, Z = Mean of actual value

$j = j^{\text{th}}$ observation, R^2 = Determination coefficient

x = Actual values, y = Predicted values

k = Number of predictors, m = Number of observations

$\sum L(X,Y)$ = Total pinball loss using model's

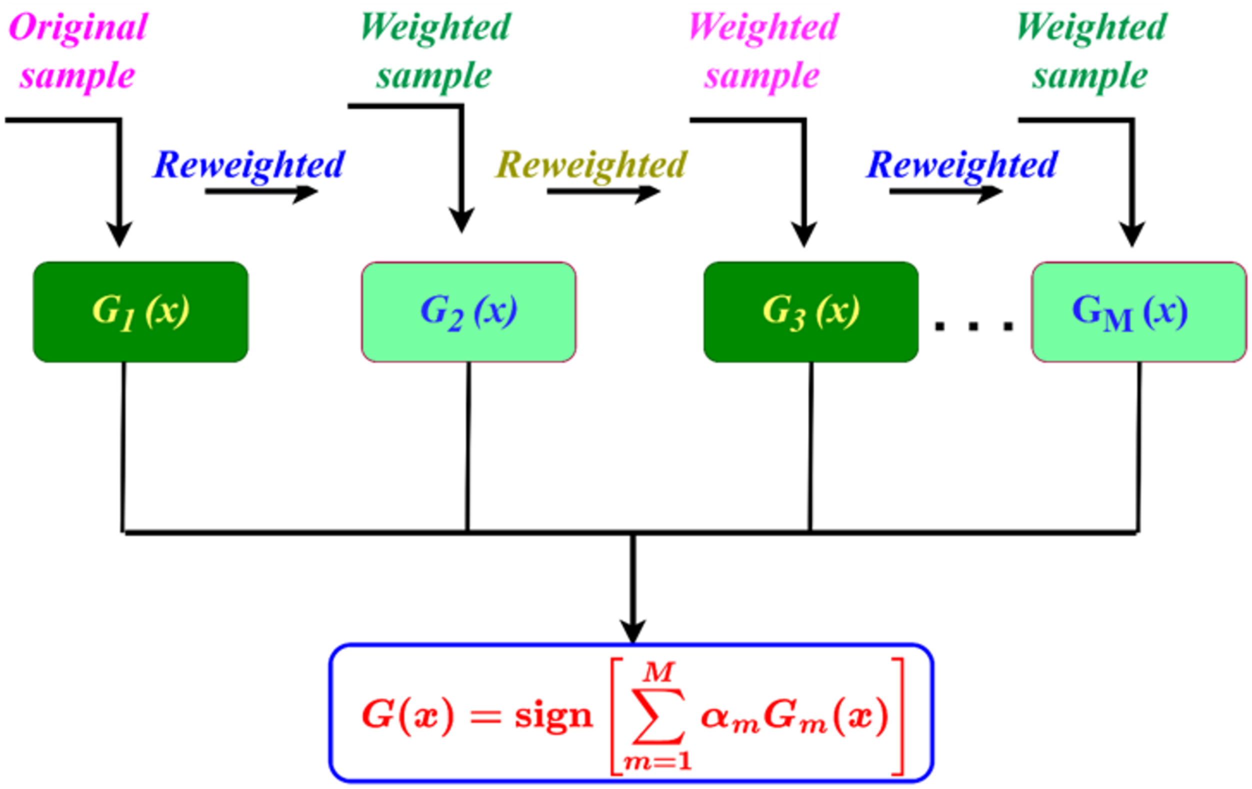

Figure 4. Setup of ADBR.

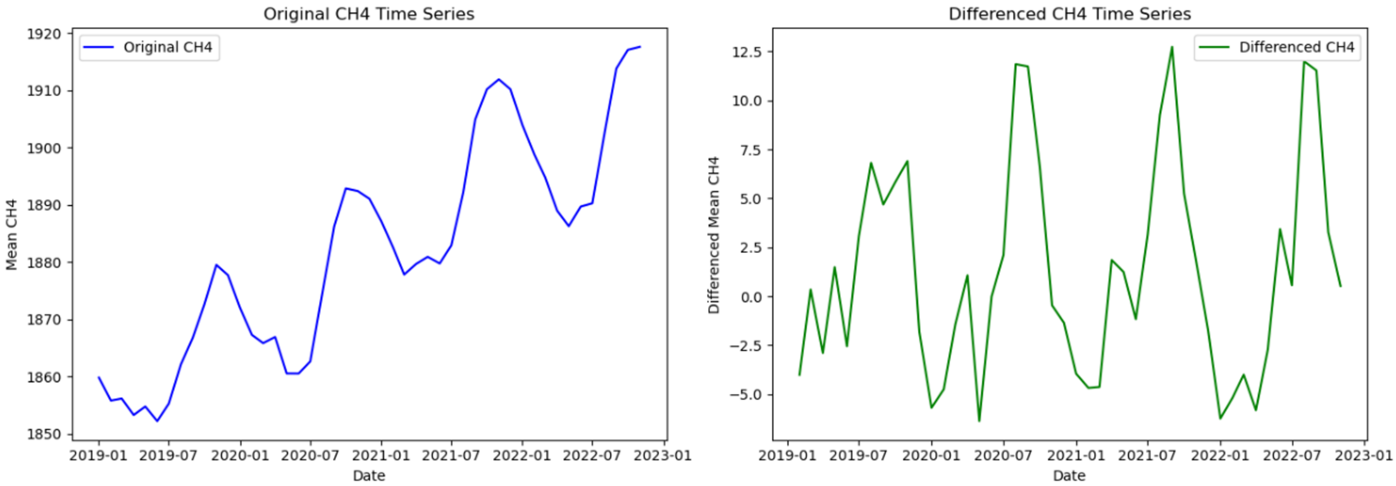

Figure 5. Non-stationary and Stationary Curve.

predictions

$\sum L(X, Z)$ = Total pinball loss using mean

D = Each observation's individual tweedie deviance

4 Results and Discussions

4.1 Comparison of ML models

Here, eight different ML models were used to predict CH_4 emission at global level. Figure 9 shows the actual values of the first five rows of mean_ch4, as well as the predicted values with their corresponding prediction errors. For example, the actual value was 1841.82, but its predicted value is 1837.20, resulting in a mistake of -4.62. The 'mean_ch4_anomaly' variable of the methane emission dataset (ADBR and XGBR) had the highest impact on mean_ch4 (target variable), while DTR and LGBR had the lowest impact on the same variable. The essential variables in almost all models, such as 'mean_ch4_anomaly', have a greater influence on the target variable than other variables in the methane emission dataset. Figure 10 shows the prediction of regression and boosting models. Random Forest Regressor (RFR) performs the best among these models, followed by CatBoost Regressor

Table 5. ML Model's performance.

Error Metrics	Model's Name							
	LR	XGBR	LGBR	CABR	DTR	RFR	ADBR	GBR
MSE	297.6791	142.9567	161.9729	141.8725	229.8335	126.0808	507.0289	190.0142
RMSE	17.2534	11.9564	12.7269	11.9110	15.1603	11.2286	22.5173	13.7846
MAE	12.1566	7.4822	8.0084	7.4597	8.1886	6.6712	19.4860	8.7469
MAPE	0.0065	0.0040	0.0043	0.0040	0.0044	0.0036	0.0103	0.0047
R^2 Score	0.6891	**0.8497**	**0.8297**	**0.8509**	0.7584	**0.8675**	0.4670	0.8003
Adjusted R^2 Score	0.6891	0.8497	0.8297	0.8509	0.7584	0.8675	0.4670	0.8003
EVS	0.6871	0.8497	0.8297	0.8509	0.7584	0.8675	0.6809	0.8003
MeAE	8.8334	4.7842	5.2478	4.7523	3.8152	3.8819	19.0092	5.8327
ME	98.3649	348.8019	346.6886	346.2459	270.2844	268.1095	299.9635	314.0386
MSLE	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001	0.0001
D2AES	0.4881	0.6849	0.6628	0.6859	0.6552	0.7191	0.1795	0.6317
D2PS	0.4881	0.6849	0.6628	0.6859	0.6552	0.7191	0.1795	0.6317
D2TS	0.6839	0.8475	0.8276	0.8490	0.7544	0.8657	0.4636	0.7974
MPD	0.1587	0.0763	0.0865	0.0758	0.1231	0.0674	0.2697	0.1016
MGD	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001	0.0001
MTD	0.0037	0.0018	0.0020	0.0018	0.0028	0.0016	0.0062	0.0023

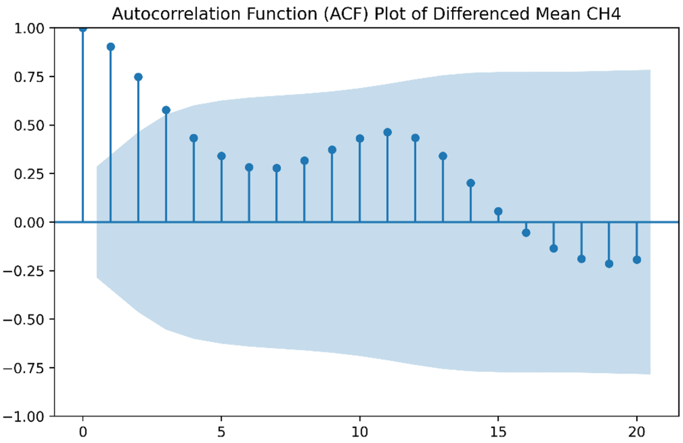

Figure 6. ACF Curve.

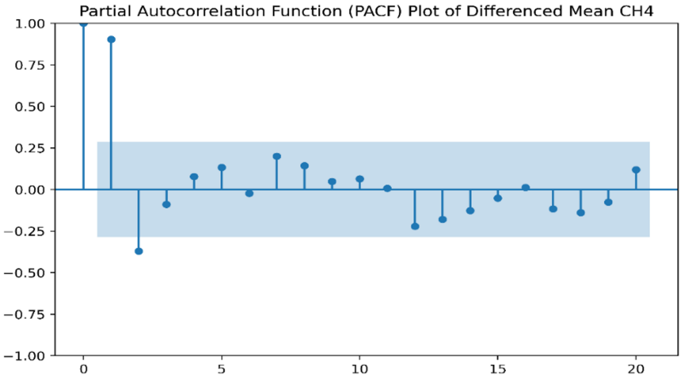

Figure 7. PACF curve.

Table 6. Forecasted values of TS models.

Model	Year Month	Forecasted value
ARIMA	2023-01-01	1910.83
	2023-02-01	1904.17
	2023-03-01	1899.56
	....	
	2042-09-01	1905.67
	2042-10-01	1905.67
	2042-11-01	1905.67
SARIMA	2023-01-01	1910.24
	2023-02-01	1905.03
	2023-03-01	1900.51
	....	
	2042-09-01	1950.82
	2042-10-01	1953.95
	2042-11-01	1954.12
LSTM	2023-01-31	1914.68
	2023-02-28	1915.60
	2023-03-30	1917.03
	....	
	2042-09-30	1957.71
	2042-10-31	1957.71
	2042-11-30	1957.71

(CABR) and XGBoost Regressor (XGBR). Figure 11 shows the actual and predicted values of tree-based models. Gradient boosting accuracy is higher than that of different models, such as AdaBoost, decision trees, and random forests.

Table 5 shows the performance metrics of various ML models used to predict the global CH_4 emissions from 2019 to 2022. Using actual CH_4 emissions data, we assessed how well these models predicted CH_4 emissions in the real world. The predicting accuracy

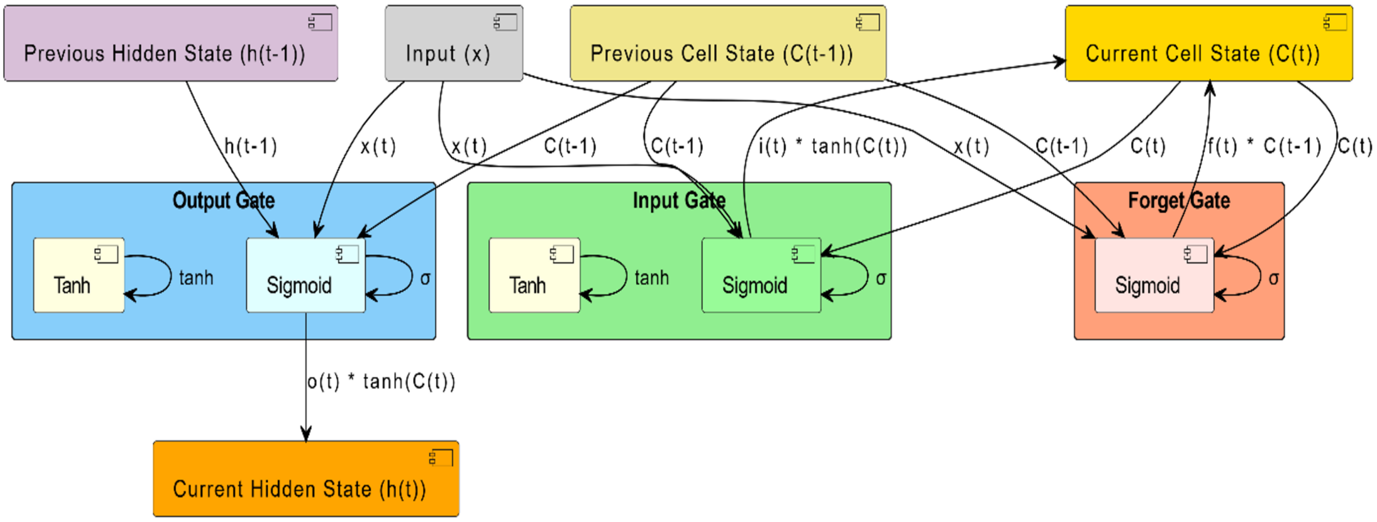

Figure 8. PACF curve.

Table 7. Error metrics comparison of TS and top four ML models.

Error Metrics	Model's Name						
	ARIMA	SARIMA	LSTM	RFR	CABR	XGBR	LGBR
MSE	17.3248	7.9126	379.9713	126.0808	141.8725	142.9567	161.9729
RMSE	4.1623	2.8129	19.4928	11.2286	11.9110	11.9564	12.7269
MAE	3.0002	1.8391	15.0276	6.6712	7.4597	7.4822	8.0084
MAPE	0.0016	0.0010	0.0079	0.0036	0.0040	0.0040	0.0043
R^2 Score	0.8704	0.9408	-1.3221	0.8675	0.8509	0.8497	0.8297

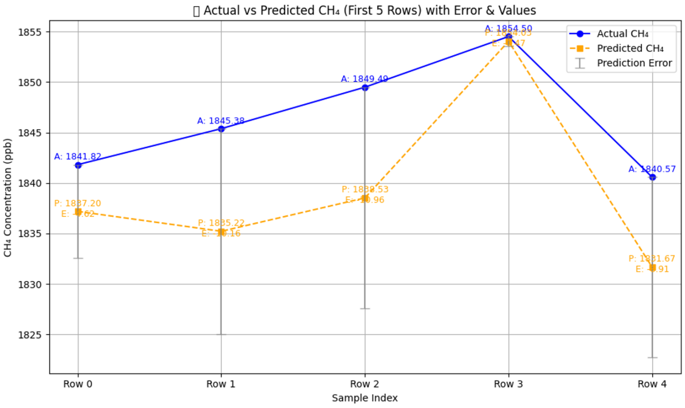

Figure 9. Actual v/s Predicted value.

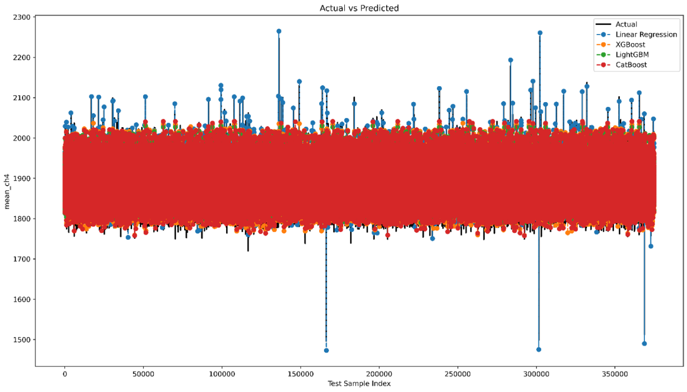

Figure 10. Actual v/s Predicted value of LR, XGBR, LGBR, and CABR.

of the RFR model was higher than that of any other ML models. The RFR predicted methane emissions with the lowest RMSE (11.2286), MAE (6.6712), MSE (126.0808), MAPE (0.0036), and MTD (0.0016) values, and the almost maximum R² (0.8675), Adjusted R² (0.8675), D2AES (0.7191), D2PS (0.7191), D2TS (0.8657), and EVS (0.8675) values. Additionally, the XGBR model demonstrated remarkable performance, even during the forecasting phase. Similarly, the ABDR model predicted methane emissions with the lowest R^2 (0.4670), adjusted R^2 (0.4670) value, and the highest MSE (507.0289), RMSE (22.5173), and MAE (19.4860) scores.

4.2 Time series model's comparison

In this study, three time series models ARIMA, SARIMA, and LSTM were used to predict the emission of CH_4 gas at the international level. Figures 12, 13, and 14 display the observed and forecasted values of CH_4 using SARIMA, ARIMA, and LSTM models, along with their corresponding confidence levels. The blue line shows the actual value of Methane emission

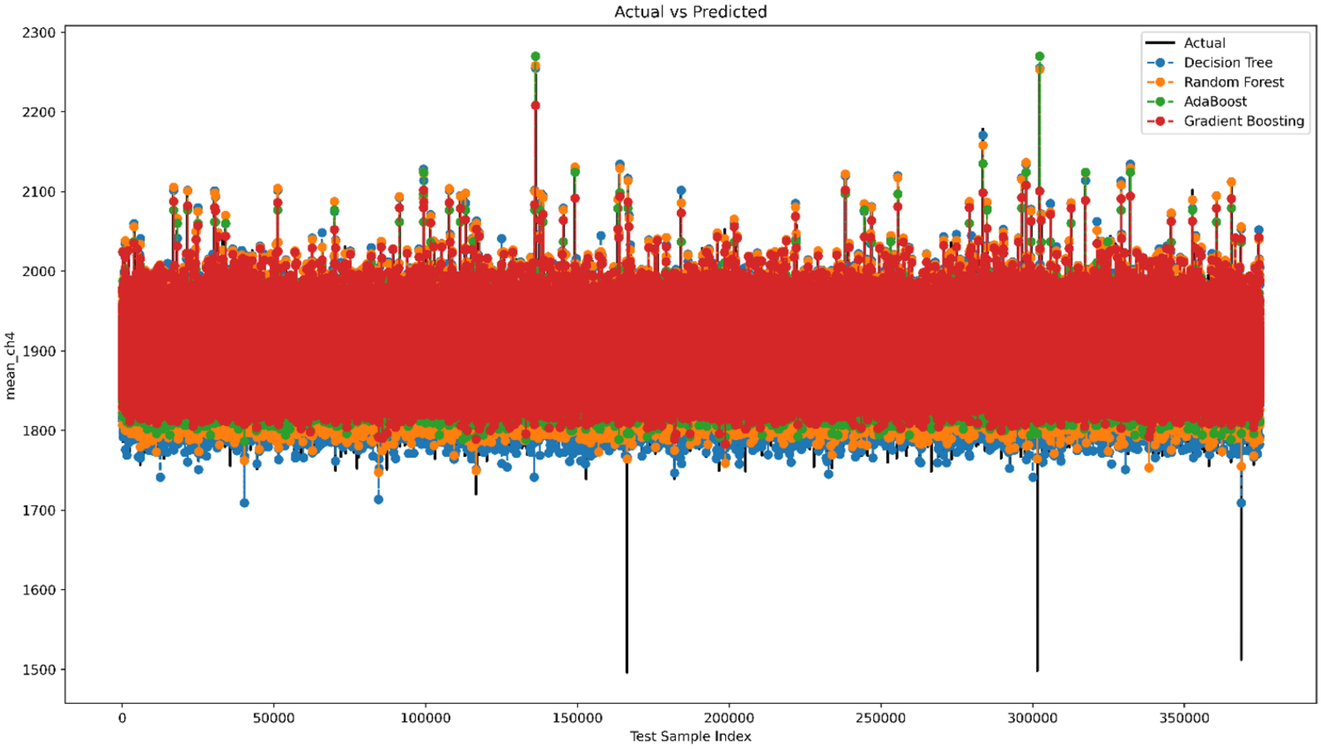

Figure 11. Actual v/s Predicted value of DTR, RFR, ADBR, and GBR.

data, whereas the green and orange lines show the forecast value of CH_4. The shaded region indicates the confidence level of models, which represents the uncertainty bounds of the Methane forecast. The lower and upper values of confidence intervals for mean_ch4 for 2023-01-01 are 1903.22 (lower) and 1917.26 (upper). Similarly, we can find the value of the confidence interval for other months and years using ARIMA and LSTM models. Table 6 shows the forecasted value of total CH_4 emission at global level using ARIMA, SARIMA, and LSTM models.

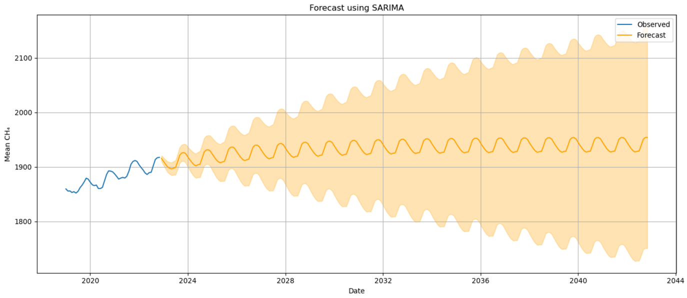

Figure 12. SARIMA model forecast and observed value.

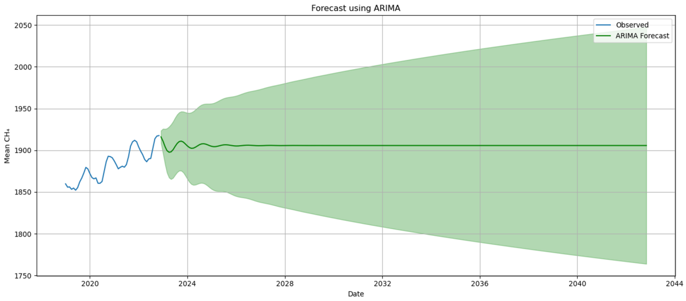

Figure 13. ARIMA model forecast value.

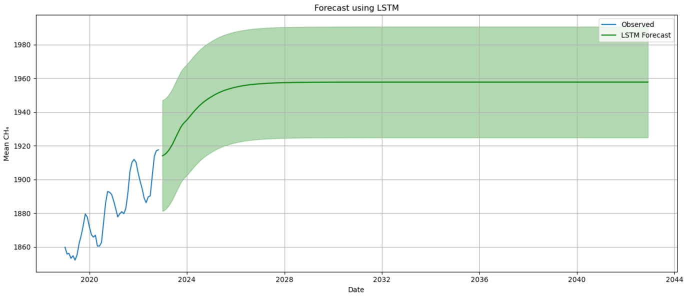

Figure 14. LSTM model forecast value.

Table 7 presents the best-performing models of TS and ML, along with their respective error metric values. The SARIMA model shows the highest R^2 squared error (94%), lowest MAPE (0.0010), MAE (1.8391), RMSE (2.8129), and MSE (7.9126), whereas LSTM shows the lowest R^2 squared (-1.3221) and the highest MSE, RMSE, MAE, and MAPE values.

5 Conclusion

This research aimed to apply ML and DL techniques to predict methane gas emissions, which have a more significant impact on global warming than CO_2 emissions. In this study, eight machine learning methods and three time series techniques were employed to identify the most effective approach. The goal was to provide initial insights for researchers to demonstrate the effectiveness of the ML and ITS methods, enabling them to apply this model later. Forecasting CH_4 emissions at an initial phase is essential. Our research indicates that the SARIMA approach surpasses other models in predicting CH_4 emissions resulting from urban activities from 2019 to 2022. The SARIMA model forecasted CH_4 emissions with the smallest RMSE, MAE, MSE, MAPE, and MTD scores, along with the greatest R^2, Adjusted R^2, D2AES, D2PS, D2TS, and EVS statistics. The created model demonstrates excellent capability in forecasting CH_4 emissions when compared to alternative ML model.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mar, K. A., Unger, C., Walderdorff, L., & Butler, T. (2022). Beyond CO2 equivalence: The impacts of methane on climate, ecosystems, and health. *Environmental Science & Policy, 134*, 127–136. [CrossRef]

[2] Allen, R. J., Zhao, X., Randles, C. A., Kramer, R. J., Samset, B. H., & Smith, C. J. (2023). Surface warming and wetting due to methane's long-wave radiative effects muted by short-wave absorption. *Nature Geoscience, 16*(4), 314–320. [CrossRef]

[3] Ahmed, T., Mahmood, Y., Yodo, N., & Huang, Y. (2024). Weather-related combined effect on failure propagation and maintenance procedures towards sustainable gas pipeline infrastructure. *Sustainability, 16*(13), 5789. [CrossRef]

[4] Magazzino, C., Madaleno, M., Waqas, M., & Leogrande, A. (2024). Exploring the determinants of methane emissions from a worldwide perspective using panel data and machine learning analyses. *Environmental Pollution, 348*, 123807. [CrossRef]

[5] Awulu, E. (2021). Gas flaring in Nigeria: A crisis for the environment. *Academia Letters*, 1–6. [CrossRef]

[6] Ezinna, P. C., Ugwuibe, C. O., & Okwueze, F. O. (2024). Gas flaring, sustainable development goal 2, and food security reflections in the Niger Delta area of Nigeria. *Discover Global Society, 2*(1), 61. [CrossRef]

[7] Stern, J. P. (2025). *Measurement reporting and verification of methane emissions from the gas and oil sector and consequences for LNG trade: A three year progress report* (No. 46). OIES Paper: ET. [CrossRef]

[8] Zhang, B., LI, H., Zhong, B., & GAO, J. (2022). The situation, problems and countermeasures for the controls of China's methane emissions. *China Mining Magazine, 31*(2), 1-10. [CrossRef]

[9] Khanna, N., Lin, J., Liu, X., & Wang, W. (2024). An assessment of China's methane mitigation potential and costs and uncertainties through 2060. *Nature Communications, 15*(1), 9694. [CrossRef]

[10] Mousavi, S. M., Dinan, N. M., Ansarifard, S., Borhani, F., Darvishi, A., Mustafa, F., & Naghibi, A. (2024). Unveiling the drivers of atmospheric methane variability in Iran: A 20-year exploration using spatiotemporal modeling and machine learning. *Environmental Challenges, 15*, 100946. [CrossRef]

[11] Jahan, I., Mehana, M., Matheou, G., & Viswanathan, H. (2024). Deep Learning-Based quantifications of methane emissions with field applications. *International Journal of Applied Earth Observation and Geoinformation, 132*, 104018. [CrossRef]

[12] Ehinmowo, A. B., Nwaneri, B. I., & Olaide, J. O. (2025). Predictive modeling of hydrogen production and methane conversion from biomass-derived methane using machine learning and optimisation techniques. *Next Energy, 7*, 100229. [CrossRef]

[13] Luo, R., Wang, J., & Gates, I. (2023). Machine learning for accurate methane concentration predictions: Short-term training, long-term results. *Environmental Research Communications, 5*(8), 081003. [CrossRef]

[14] Chen, S., Liu, L., Ma, Y., Zhuang, Q., & Shurpali, N. J. (2024). Quantifying global wetland methane emissions with in situ methane flux data and machine learning approaches. *Earth's Future, 12*(11), e2023EF004330. [CrossRef]

[15] Cusworth, D. H., Duren, R. M., Thorpe, A. K., Pandey, S., Maasakkers, J. D., Aben, I., ... & Miller, C. E. (2021). Multisatellite imaging of a gas well blowout enables quantification of total methane emissions. *Geophysical Research Letters, 48*(2), e2020GL090864. [CrossRef]

[16] Wang, Z., Feng, M., Johnson, M. F., Lipani, A., & Chan, F. (2025). The role of reservoir size in driving methane emissions in China. *Water Research, 279*, 123441. [CrossRef]

[17] Tayarani-N, M.-H., & Paykani, A. (2025). An ensemble learning algorithm for optimization of spark ignition engine performance fuelled with methane/hydrogen blends. *Applied Soft Computing, 168*, 112468. [CrossRef]

[18] Mansouri, T. S., Wang, H., Mariotti, D., & Maguire, P. (2022). Methane detection to 1 ppm using machine learning analysis of atmospheric pressure plasma optical emission spectra. *Journal of Physics D: Applied Physics, 55*(22), 225205. [CrossRef]

[19] Luo, R., Wang, J., & Gates, I. (2024). Estimating air methane and total hydrocarbon concentrations in Alberta, Canada using machine learning. *Atmospheric Pollution Research, 15*(2), 101984. [CrossRef]

[20] Wójcik-Gront, E., & Wnuk, A. (2025). Evaluating methane emission estimates from Intergovernmental Panel on Climate Change compared to Sentinel-derived air-methane data. *Sustainability, 17*(3), 850. [CrossRef]

[21] Li, K., Bai, K., Jiao, P., Chen, H., He, H., Shao, L., ... & Chang, N. B. (2024). Developing unbiased estimation of atmospheric methane via machine learning and multiobjective programming based on TROPOMI and GOSAT data. *Remote Sensing of Environment, 304*, 114039. [CrossRef]

[22] Chang, H. T., Chern, Y. R., Asri, A. K., Liu, W. Y., Hsu, C. Y., Hsiao, T. C., ... & Wu, C. D. (2025). Innovating Taiwan's greenhouse gas estimation: A case study of atmospheric methane using GeoAI-Based ensemble mixed spatial prediction model. *Journal of Environmental Management, 380*, 125110. [CrossRef]

[23] Wei, X., Huang, W., Liu, L., Wang, J., Cui, Z., & Xue, L. (2024). Low-rank coalbed methane production capacity prediction method based on time-series deep learning. *Energy, 311*, 133247. [CrossRef]

[24] Wang, D., Li, Z., & Fu, Y. (2024). Production Forecast of Deep-Coalbed-Methane Wells Based on Long Short-Term Memory and Bayesian Optimization. *SPE J, 29*(7), 3651-3672. [CrossRef]

[25] Ai, X., Hu, C., Yang, Y., Zhang, L., Liu, H., Zhang, J., ... & Xiao, W. (2024). Quantification of Central and Eastern China's atmospheric CH4 enhancement changes and its contributions based on machine learning approach. *journal of environmental sciences, 138*, 236-248. [CrossRef]

[26] Gu, X., Yao, L., Xiao, X., Wu, L., & Yu, H. (2025). Optimizing methane flux prediction and key feature identification based on a novel hybrid machine learning model. *Iscience, 28*(12). [CrossRef]

[27] Venkatesa, P. N. B., Kalpana, M., Balakrishnan, N., Balamurugan, V., Suresh, A., Rajavel, M., & Dhivya, R. (2025). Using machine learning models to forecast methane emissions from agriculture in India. *Plant Science Today, 12*(2). [CrossRef]

[28] Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., aan de Brugh, J., Schneider, A., ... & Landgraf, J. (2021). Methane retrieved from TROPOMI: improvement of the data product and validation of the first 2 years of measurements. *Atmospheric Measurement Techniques, 14*(1), 665-684. [CrossRef]

[29] Suri, N. M. R., Murty, M. N., & Athithan, G. (2019). *Outlier detection: techniques and applications*. Springer Nature. [CrossRef]

Suraj Arya currently working as Assistant Professor in the Department of Computer Science and Information Technology and Deputy Director (Training and Placement) at Central University of Haryana, India. His academic qualifications are Ph.D.(Computer Science), M.Phil.(Computer Science) and M.Tech (Computer Science and Engineering). His research interests focus on machine learning (ML), Deep Learning, internet of things (IoT), Data warehousing and mining, system automation and patents writings. He has granted and files many patents. He has also published many research articles in international journals, book chapters, and conferences. (Email: surajarya@cuh.ac.in)

Anju is a research scholar of Central University of Haryana, India. She received her B.Tech. in Computer Science and Engineering from Maharshi Dayanand University, Rohtak and M.Sc. in Computer Science from Chaudhary Bansi Lal University, Bhiwani. She is currently doing her Ph.D. (Computer Science) from Central University of Haryana. Her research interests focuses on Machine Learning (ML), and Internet of Things (IoT). (Email: anju24sanga@gmail.com)

Jonas Nnaemeka Onah stands out as a prominent academic and expert in Electrical Engineering, focusing on Power Systems, High Voltage, Renewable energy, and Artificial Intelligence. Based at the Federal University of Petroleum Resources Effurun, Delta State, Nigeria, he actively imparts knowledge and drives research. With a prolific publication record in both local and international journals, Dr. Onah showcases his research fervor. Notably, he accomplished his Ph.D at the University of Nigeria, Nsukka, in an impressively short timeframe within the Electrical Engineering department, underscoring his academic prowess and contributions that cement his stature in the field. (Email: onah.jonas@fupre.edu.ng)