PLANT SCIENCE TODAY
Vol 12(2): 1-12
https://doi.org/10.14719/pst.7010

eISSN 2348-1900

RESEARCH ARTICLE

Using machine learning models to forecast methane emissions from agriculture in India

Venkatesa Palanichamy Narasimma Bharathi[1], Kalpana Muthuswamy[1*], Balakrishnan Natarajan[1], Balamurugan Vasudevan[1], Suresh Appavu[1], Rajavel Marimuthu[2] & Dhivya Rajaram[3]

[1]Department of Agriculture, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[2]Office of the Public Relations, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[3]Department of Business Administration, PSGR Krishnammal College for Women, Coimbatore 641001, Tamil Nadu, India

*Correspondence email - kalpusiva@gmail.com

Received: 03 March 2024; Accepted: 10 March 2025; Available online: Version 1.0: 30 April 2025; Version 2.0 : 24 June 2025

Cite this article: Venkatesa PNB, Kalpana M, Balakrishnan N, Balamurugan V, Suresh A, Rajavel M, Dhivya R. Using machine learning models to forecast methane mmissions from agriculture in India. Plant Science Today. 2025; 12(2): 1-12. https:/doi.org/10.14719/pst.7010

Abstract

Methane (CH_4) is a potent and powerful greenhouse gas with significant warming potential. CH_4 has approximately 84 times the global warming potential (GWP) of CO_2 over a 20-year period and 25 times over a 100-year period. Methane persists in the atmosphere for approximately a decade before breaking down through oxidation processes. In order to forecast methane emissions from agricultural and related activities, this study applied an evaluation approach. This study analyzed annual data from 1990 to 2021, collected from the FAOSTAT website, using 7 machine learning models: Random Forest, LASSO, Gradient Boosting, AdaBoost, XGB, Ridge Regression and Linear Regression. The results indicate that the linear regression model outperformed the other models in predicting methane emissions. The results show that linear regression is more effective than the various machine learning algorithms. The linear regression model (R^2 = 0.98, RMSE = 1.95, MSE = 3.86 and MAE = 1.45) achieved the best performance among all models in terms of accuracy. The comparative study's findings should yield a highly accurate assessment of the methane emissions from agricultural areas and help with the creation of laws or other policies aimed at reducing those emissions. These findings can assist the Indian government in formulating effective policies to mitigate methane emissions while maintaining agricultural productivity. Our approach estimates methane emissions from agriculture in India with an R^2 value of 0.99, indicating high predictive accuracy.

Keywords: agriculture; forecasting; machine learning; methane emission

Introduction

India has a significant agricultural sector that plays a major role in the nation's economic development. A sustainable, secure and profitable agricultural system is essential for long-term food security. The agricultural sector in India is expanding due to increasing food demand, technological advancements and dietary shifts. Farmers have adopted various methods to enhance agricultural productivity to meet the raising food demand (1). While food production increases to meet demand, agricultural activities contribute to greenhouse gas emissions, particularly releases methane from rice cultivation and livestock and carbon dioxide (CO_2) from machinery and land-use changes. Methane from agriculture is the primary contributor to the increase in greenhouse gas emissions (2). Agricultural operations are the primary contributors to greenhouse gas emissions within the food systems. Megatonnes of CH_4 are released into the atmosphere annually by the agriculture industry, which has detrimental impacts and releases additional pollutants. Livestock enteric fermentation and rice cultivation were the primary sources of methane emissions (Press Information Bureau, 2023). Although large-scale manufacturing makes sense, environmental issues are vital to human welfare. More than 100 countries pledged to cut global methane emissions by 30 % by the end of the decade compared to 2020 levels during the UN Climate Change Conference in Glasgow. Experts estimate that global warming could be reduced by 0.2 °C by 2050 if the commitments made at COP26 (2021 United Nations Climate Change Conference) are fully implemented. Along with China and Russia, the two biggest emitters of methane, India, the third largest in the world, will be noticeably absent from the list of contributors (Fig. 1). Scientists in India caution that while the need to cut methane emissions worldwide is mounting, doing so would necessitate a dramatic overhaul of the country's agricultural system, which may not be technically and economically possible. In 2016, India's methane emissions (excluding LULUCF) accounted for 409 million tonnes of CO_2 eq. Agriculture accounted for 74 % of methane emission, followed by the waste sector 14.5 %, the energy sector, 10.6 % and the industrial process and product use 1 %. Thus, agriculture remains the dominant source of methane emissions in India (3).

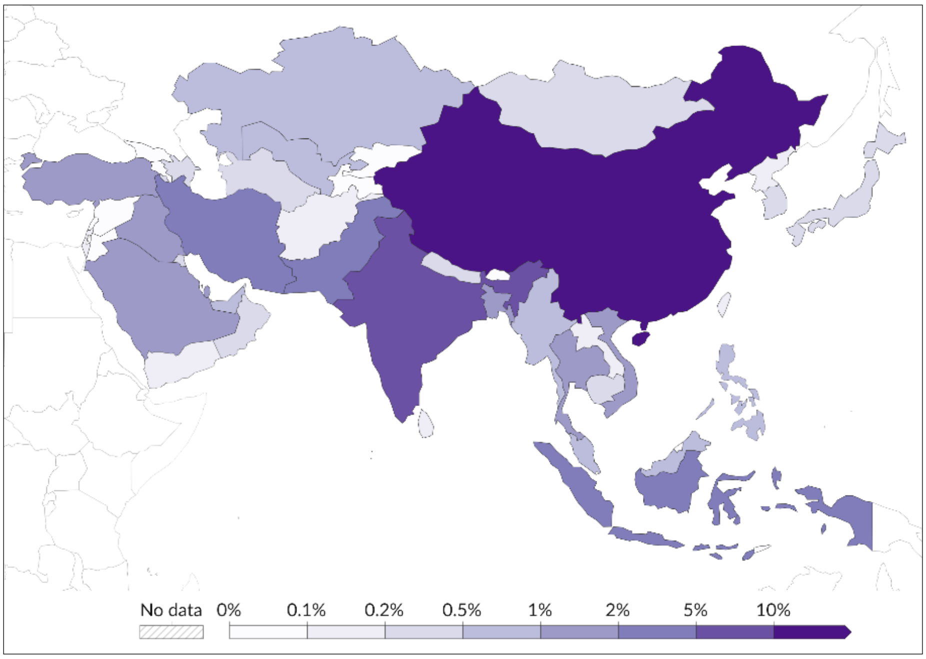

Fig. 1. Share of Asian countries methane emissions from fossil fuels, industry and agricultural sources, 2021.

Data source: OurWorldInData.org/co₂-and-greenhouse-gas-emissions (32).

Methane is a hydrocarbon gas consisting of 4 hydrogen atoms and 1 carbon atom. Methane is one of the most potent greenhouse gases due to its high global warming potential (GWP). Methane emissions arise from both natural sources (wetlands, lakes, geological activity) and anthropogenic activities (agriculture, fossil fuel extraction and waste management). Over the past century, human activity has caused the atmospheric concentration of methane to double (4). Methane is the second most significant greenhouse gas after carbon dioxide, accounting for 23 % global radiative forcing (5). Unlike carbon dioxide, which can persist in the atmosphere for centuries, methane has a shorter atmospheric lifetime of approximately 12 years. However, over a 20-year period, methane is about 80 times more effective at trapping heat than CO_2. Methane can further break down into carbon dioxide when it combines with other atmospheric gases, such as oxygen. The monthly global average of atmospheric methane concentration reached a high concentration of 1924.65 parts per billion (ppb) in December 2022 (6). The global average yearly concentration of atmospheric methane has increased significantly during the last 30 years (Fig. 2). Since methane gas emissions have the potential to cause global warming 27 times more than carbon dioxide, swift action is needed to reduce them to the objective of 1.5 °C by 2050 (7). According to a study, anthropogenic sources of methane include waste management, agriculture, mining and fossil fuels combustion, while natural sources include wetlands, lakes, streams and geological activity (4). One of the primary causes of climate change is the increase in natural greenhouse gas emissions, which cannot be stopped because human activity will always produce gases like CO_2, CH_4 and nitrous oxide (N_2O). To limit global warming to 1.5 °C, greenhouse gas emissions must be reduced by 45 % form 2010 levels by 2030 (8). It was decided that between 2022 and 2023, this rate would be 1 %. 52 out of 198 nations have not reported a notable increase in methane gas emissions under the UN Framework Convention on Climate Change (COP-27, Assessment Reports) (9). It was discovered that there is a large interest in research on greenhouse gas emissions, especially carbon dioxide emissions. Research on methane gas, one of the main causes of global warming, has been less

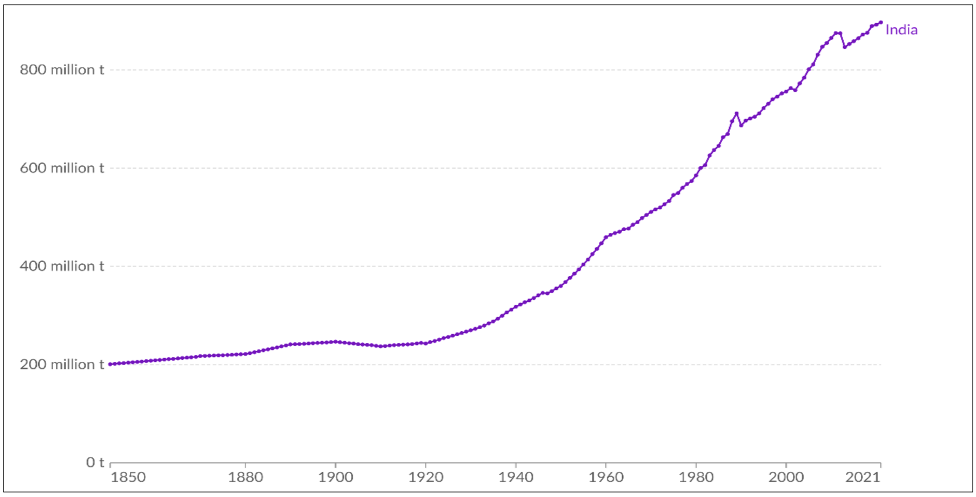

Fig. 2. Decadal variations in the amount of CH_4 released into the Indian atmosphere between 1850 and 2021 from sources such as industry, agriculture and fossil fuels. Tonnes of CO_2 eq are used to quantify CH_4 emissions (32, 33).

extensive. The study's prediction of methane gas will aid nations in creating emission reduction plans. It is also expected to act as a guide for India's attempts to reduce emissions at COP 28, which is scheduled for November 30, 2023, in Dubai. The study is expected to contribute to the body of knowledge on methane gas emissions from agricultural operations in India, as there has been a lot of research on this topic.

Materials and Methods

This study used 7 distinct machine learning techniques- Random Forest (RF), LASSO, Gradient Boosting, Ada boost, XGB, Ridge Regression and Linear Regression to forecast methane gas emissions from agricultural operations in India. The dependent variable was the total amount of methane gas emitted and forecasting algorithms were created using eight independent factors. To increase the efficacy rate of the investigation, cross-validation was used. The statistical measures R^2, MAE, RMSE and MSE were used to evaluate the models' accuracy and success.

Data collection and preparation

For data analysis to be successful, access to pertinent data is essential. The study utlilzed data from FAOSTAT and Climate Watch records covering the period from1990 to 2021 (10, 11). Missing data were estimated using interpolation. The independent variables include enteric fermentation, rice production, crop residue burning, agricultural and food waste, manure management, pesticide production, on-farm energy use and household food consumption. The total methane emissions served as the dependent variable.

Machine learning

Machine learning is a computational approach that enables systems to learn patterns from data and make predictions or classifications. The scientific field of machine learning examines attempts to develop various models, algorithms and learning techniques that enable machines to acquire knowledge in a manner akin to that of human beings. By using statistical and mathematical techniques to data and deriving inferences from forecasts, it discusses instructional tactics and their efficacy (12).

Linear regression

The prediction was tested using the most basic linear regression modelling approach among machine learning techniques. The approach of multiple linear regressions is applied based on a set of independent variables. When compared to alternative algorithms, linear regression has comparable performance, learns quickly and has a strong explanatory power. A linear relationship between one or more independent variables and a dependent variable is simulated by a representative regression technique known as linear regression (13). The link between the independent and dependent variables is explained by the linear expression (14, 15).

This formula is expressed as

$$y = w_1x_1 + w_2x_2 + w_3x_3 + \ldots\ldots + w_8x_8 + b \qquad \text{(Eqn. 1)}$$

y is the dependent variable (output value), $x_1, x_2, \ldots\ldots x_8$ is the independent variables (input value), $w_1, w_2, \ldots\ w_8$ is the independent variables weight and b is the bias.

Random Forest (RF)

The Random forest algorithm is an efficient and versatile data mining technique method that performs well with high-dimensional data. The process is predicated on building several regression or classification trees. According to a study, only RF uses this subset of features to build a single tree by using a bootstrap sample and random selection of an array of the variables (16). Bootstrap samples are taken after variables are selected at random multiple times and the results are aggregated (17). An ensemble model is a model that uses model combining to boost prediction probability. Put another way, the ensemble model performs better in terms of predictive power than other single models. As Fig. 3 shows, the RF prediction strategy uses regression trees to generate an ideal model. RF employs bagging (Bootstrap Aggregating), a technique where multiple regression trees are trained on bootstrapped subsets of data and their predictions are aggregated to improve model accuracy and reduce overfitting (18). After bagging is used to construct Tree -1, the process is repeated multiple times, with each tree trained on a different bootstrap sample. Tree-1, Tree-2 and up to Tree-100 are constructed using this iterative process. A forecast is created by taking the trees' average. RF regression is widely used in predictive modelling and numerical analysis, as demonstrated in (19).

Gradient Boost Regression (GBR)

The proposed gradient boosting method, is a widely used machine learning algorithm known for its effectiveness (20). Weak learners can be combined to create strong learners using the gradient boosting method. This method bases the classification on the residuals from the preceding iteration, evaluating the impact of each feature one at a time until the target accuracy is attained (Fig. 4). The Loss function L(φ), which calculates the residuals, is optimized by gradient descent (21). The sum of the outputs of the K successive classifier functions f_k yields the final result φ(X) as follows:

$$\hat{Y} = \Phi(X) = \sum_{k=1}^{K} f_k(X) \quad f_k \in F \qquad \text{(Eqn. 2)}$$

Where, K is the total number of iterations in the boosting method and f_k is a decision tree.

Extreme Gradient Boosting (XGB)

Gradient Boosting Machine is a powerful supervised learning algorithm and XGBoost is one of its most efficient implementations. It is used for both regression and classification tasks. Because of XGBoost's fast out-of-core compute execution, data scientists like it (22). Because of XGBoost's improved performance over other techniques especially in numerous data contests and machine learning competitions it has gained popularity among data researchers. XGBoost has demonstrated its ability to perform regression and classification across various applications, including predicting store sales, customer behavior analysis, ad click prediction, risk assessment, text classification and malware detection. In an ensemble method known as "boosting," flaws from earlier models are corrected by

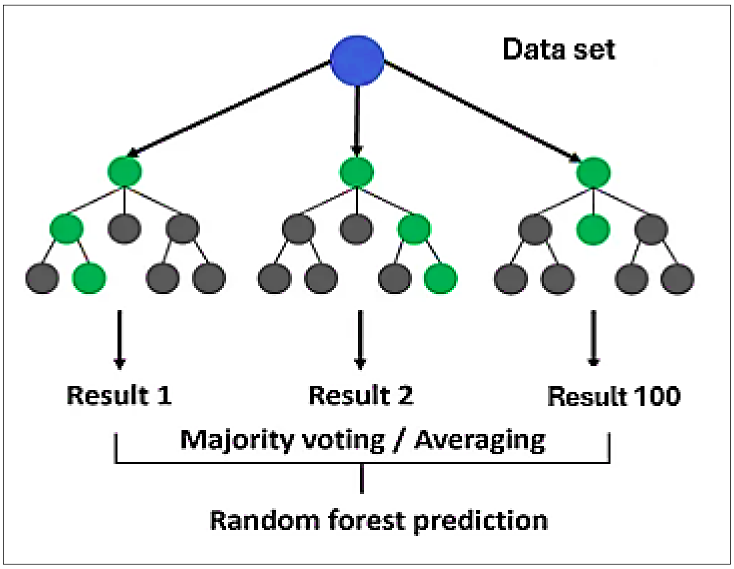

Fig. 3. Simplified structure of RF model.

appending new ones. One model is added after another until no further advancements are feasible. XGBoost employs a tree ensemble model composed of multiple Classification and Regression Trees (CART), where each successive tree corrects the errors of the previous ones through gradient boosting (20). The ensemble method is employed since a single CART usually has little predictive power. A tree ensemble model aggregates predictions from multiple CART models to improve overall accuracy.

Ada boost Regression (ABR)

Combining multiple weak classifiers to produce a single, potent classifier is known as "boosting". It is flexible in that the next classifiers that run are adjusted according to the cases that the previous classifiers misclassified. It is possible to view the strict focus on the data used for training samples that the previous weak classifier mislabeled as each weak classifier making every effort to increase the overall classification rate. AdaBoost repeatedly uses the weak classifiers to run a series of t = 1 ,..., 50 classifiers (23). The "weight" assigned to the incorrectly categorized cases increases (or decreases) with each categorization, contingent on the circumstances. New classifiers are restricted to focusing on cases that previous classifiers misclassified (24).

Ridge regression

In OLS regression analysis, multicollinearity indicated by VIF values is addressed using ridge regression. It is a useful biased estimating technique that improves the overall prediction performance of models by reducing multicollinearity. The method, which was first put forth by Hoerl and Kennard in 1970, has been shown to be better predictive when a small amount of bias is included to the model (25). This keeps overfitting from happening and lowers overall variance.

This feature makes it a good fit for estimating machine learning models, especially since MLM explanatory variables typically show multicollinearity (26). The following matrix formula (Standardized cost model- Ridge Regression) is used to $J(\theta) = \Sigma_{i=,1}^{n} (yi - \hat{y}i)^2/2n + \alpha\ \Sigma^{mi=1}\theta^{i2}$ (Eqn. 3)

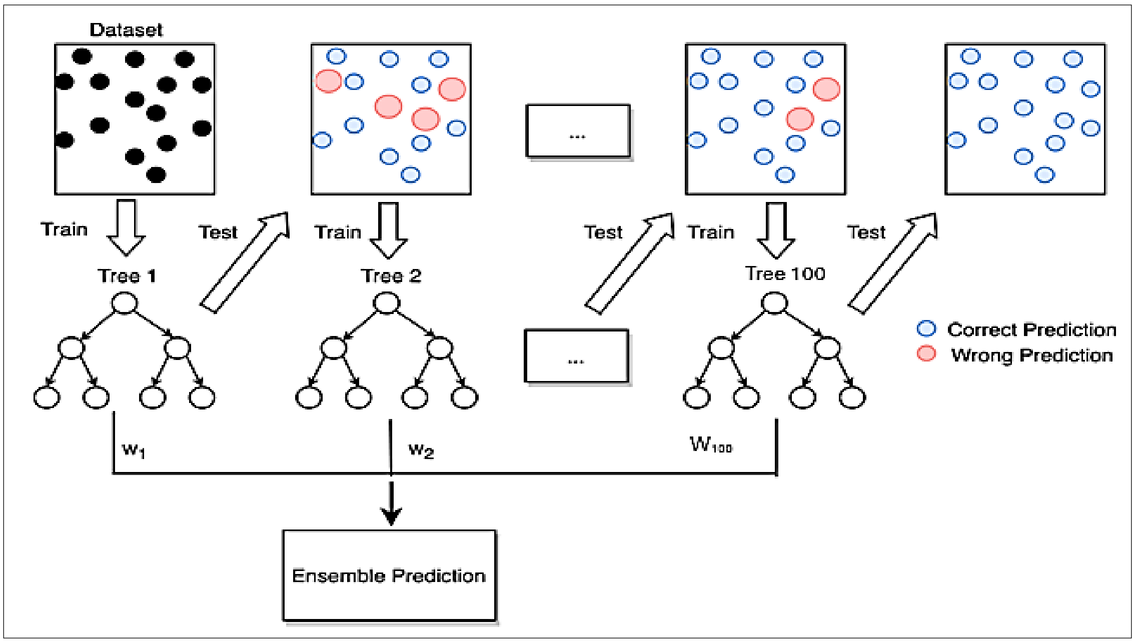

Fig. 4. Comprehensive setup of the GBR.

estimate the coefficients in ordinary least squares:

The truth value (y), the predictions (y^), the relevant parameters (θ), the penalty (α) and the loss function (J (θ) are all expressed in the above equation.

Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO is a regression analysis technique that performs both regularization and feature selection by shrinking some regression coefficients to zero, thereby eliminating less significant variables and reducing model complexity. In order to improve prediction accuracy and the understandability of regression models, this method selects a subset of the available factors that has the greatest influence on the response variable. By penalizing large regression coefficients, LASSO reduces overfitting and enhances generalization. However, if too much regularization is applied, important variables may be overly shrunk, leading to increased bias and reduced predictive accuracy. In order to make model interpretation simpler, it can also be used to select variables, removing those that greatly contribute to variation (27).

These penalized sums of squares are minimized by LASSO:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}(\beta_j x_{ij})\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad \text{(Eqn. 4)}$$

Here, X_{ij} represents the value of the j-th predictor variable for the i-th observation, 'n'(n=) denotes the number of observations and yi is the response variable for the i-th observation. The penalty parameter λ, which determines the regularization scale, shrinkage intensity and ultimately the number of variables in the final model, must be changed in order for LASSO to perform as intended. A cross-validation procedure is used to do this, with the lowest calculated prediction error yielding the recommended λ value. Although correlations between performance measures and methane emission were statistically significant, including all predictor variables in a regression model may not retain all associations as significant (17). The LASSO approach makes it easier to identify the most important performance metrics.

Tools used

Python has been chosen as the primary programming language for this study, which includes all stages from data extraction to model evaluation. An application is developed using the Python 3.6 software platform to assess the performance of the proposed design. This choice was made because Python has strong framework support in the fields of artificial intelligence and machine learning, making it perfect for real-world problem solving. Because Python is adaptable and does not require a certain operating system, the project is more accessible and flexible.

Model evaluation metrics

Cross-validation creates validation ensembles from the dataset in order to evaluate the model's performance (28). The validity and efficacy of the models were assessed using the statistical measures R^2, MAE, RMSE and MSE.

When predicting how an event will pan out, a statistical measure known as the coefficient of determination, or R^2, examines how changes in one variable may be explained by fluctuations in another. One statistical metric used to evaluate a model's predictive power is the coefficient of determination. The strength of the linear relationship between the two variables is shown by this coefficient. The outcome is represented by the dependent variable in the model. R^2 can have values as low as 0 and as high as 1. A portion of the variation in the dependent variable is explained by the model (29). The R^2 value is shown in the formula below.

$$R^2 = \frac{\left[\sum_{i=1}^{n}|y_i - \bar{y}_i|(P_i - \bar{P}_i)\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)\sum_{1=1}^{n}(P_i - \bar{P}_i)^2} \qquad \text{(Eqn. 5)}$$

The variables Y_i ,ȳ , P_i and p⊠ represent the observed and predicted greenhouse gas emission values, with ‘n’ indicating the number of samples employed in the MLMs (where i = 1, 2, ..., 14).

In statistics, the mean absolute error, or MAE, is a metric that determines the median number of deviations from a set of predictions regardless of the direction of the deviations. To calculate the percentage difference between the values that were predicted and the values that really occurred, an average across the full data set is used. MAE only measures the quantity of errors; it does not concern about their direction. A decreasing MAE improves a model's accuracy (30). The MAE formula is displayed in equation:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad \text{Eqn.6}$$

Where, ‘e’ represents the error value and ‘n’ the number of data points.

In regression models, the mean squared error (MSE) between the expected and actual values is a frequently used model evaluation metric. Since MSE squares the mistake, it highlights severe inaccuracies. The MSE is a metric that has a range of 0 to infinity; values that are closer to zero are thought to be better. A model that derives from the same set of data and has a less MSE value is considered accurate (31).

$$\mathrm{MSE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} e_i^2 \qquad \text{Eqn. 7}$$

Where, ‘e’ represents the error value and ‘n’ the number of data points.

One of the most popular evaluation standards to assess the model's truthfulness is the root mean squared error (RMSE), which is computed from:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - P_i)^2} \qquad \text{Eqn. 8}$$

In this case, 'n' indicates the total number of data points, 'Y' indicates the true methane concentration and 'P' represents the predicted value. The RMSE determines an average volume of the computation error and assigns greater weight to large errors. It's obvious that a prediction's accuracy increases with decreasing root mean square error (RMSE) values.

Results

An overview of the results obtained from using the model covered in the previous chapter is given in this part. To gain a comprehensive understanding of the dataset, we first examine the outcomes before applying the model. This section first presents descriptive statistics (such as mean, variance and standard deviation) to summarize the dataset, followed by an evaluation of the model's performance using key inferential statistics (such as R^2, RMSE and MAE) to assess predictive accuracy.

The scatter plot in Fig. 5 shows the methane emissions from various agricultural sectors and activities between 1990 - 2021. Emissions have been steadily rising as a result of intensive farming practices, widespread use of fertilizers, pesticide and manures. While the increase in methane emissions from agriculture is concerning, there are steps that may be taken to address the issue. The agricultural industry should adopt mitigating its impact on the environment by introducing more efficient irrigation systems, curbing food waste and promoting agroforestry practices. Sustainable methods and technology must be adopted by agricultural sectors to reduce the environmental impact of methane emissions. By implementing strategies such as precision cultivation, organic farming and improved manure management, the industry can significantly contribute to climate change mitigation and a more sustainable future. Investment in research and development of sustainable farming methods can help the agricultural industry lower methane emissions while improving soil quality and biodiversity.

Model construction and assessment

The dataset consisted of 320 data points on annual methane emissions, sourced from FAOSTAT and Climate Watch for the period 1990 - 2021. These data points were used to train and test the models. Eighty percent (80 %) of the dataset was used for model training, while the remaining 20 % was reserved for model evaluation. Eighty percent (80 %) of the dataset was used for model training, while the remaining 20 % was reserved for model evaluation. Table 1 presents the complete dataset.

Fig. 6 shows a funnel chart that illustrates the MAE metrics for various machine learning models on the testing dataset. The graph indicates that the XGB approach had the greatest RMSE values, while the LASSO and Ridge regression models had the lowest. In a funnel chart, the model at the top represents the model with the highest RMSE value among all the models and the model at the bottom represents the best performer with the lowest RMSE value.

Fig. 7 shows a bar diagram representing the MSE values of different machine learning models that were calculated using methane emissions data. The picture shows that the LR model performed best with a score of 3.86 and the XGB model scored lowest with an MSE value of 17.43. Higher MSE values suggest that a model is more accurate at predicting the data pattern when comparing models with different values.

Fig. 8 presents a bar chart of RMSE values for various machine learning models, derived from methane emissions data. The LR model performed best with a score of 3.86 and the XGB model scored lowest with an MSE value of 4.17 (Fig. 8). Higher MSE values suggest that a model is more accurate at predicting the data pattern when comparing models with different values.

Fig. 9 presents a radar chart of R^2 values for various machine learning models, computed using methane emissions test data. The radar's length serves as a measure of these variables' magnitude. The analysis's findings showed that the model based on LR had the highest R^2 value (0.988) and the XGB model had the lowest R^2 value (0.957).

MAE quantifies the average difference between the predicted and actual methane emission values, providing a straightforward measure of forecast error. MSE and RMSE emphasize larger errors by squaring differences, with RMSE offering error magnitude in the same units as emissions. The R^2 score indicates the proportion of variation in methane emissions explained by the model; values closer to 1 reflect more reliable forecasting.

Table 2 presents accuracy metrics for various machine learning models used to predict methane emissions from agricultural operations in India between 1991 and 2021. We evaluated the performance of these models in predicting real -world methane emissions using actual emissions data. The Linear Regression Model outperformed every other model in terms of forecasting accuracy. The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values were lowest and the R^2value was practically maximum for the methane emissions predicted by the linear regression. Furthermore, as evidenced by its best R^2 value, the XGB model performed exceptionally well even when evaluated inside the predicting period.

Fig. 10 presents a visual comparison of observed and predicted methane emissions for different models.

With the lowest RMSE (1.95), MSE (3.86) and MAE (1.45), along with the highest R^2 value (0.98), the linear regression model achieved the best forecast accuracy among all models. With the lowest RMSE (1.95), MSE (3.86) and MAE

Table 1. Details on dividing the methane emission dataset

Methane emission	Year	Total data	Test data	Train data
Eight variables (mention in previous section)	1990 - 2021 (31 years)	320	64 (20 %)	256 (80 %)

Forecasting upcoming years - next 10 years (2021 - 2030)

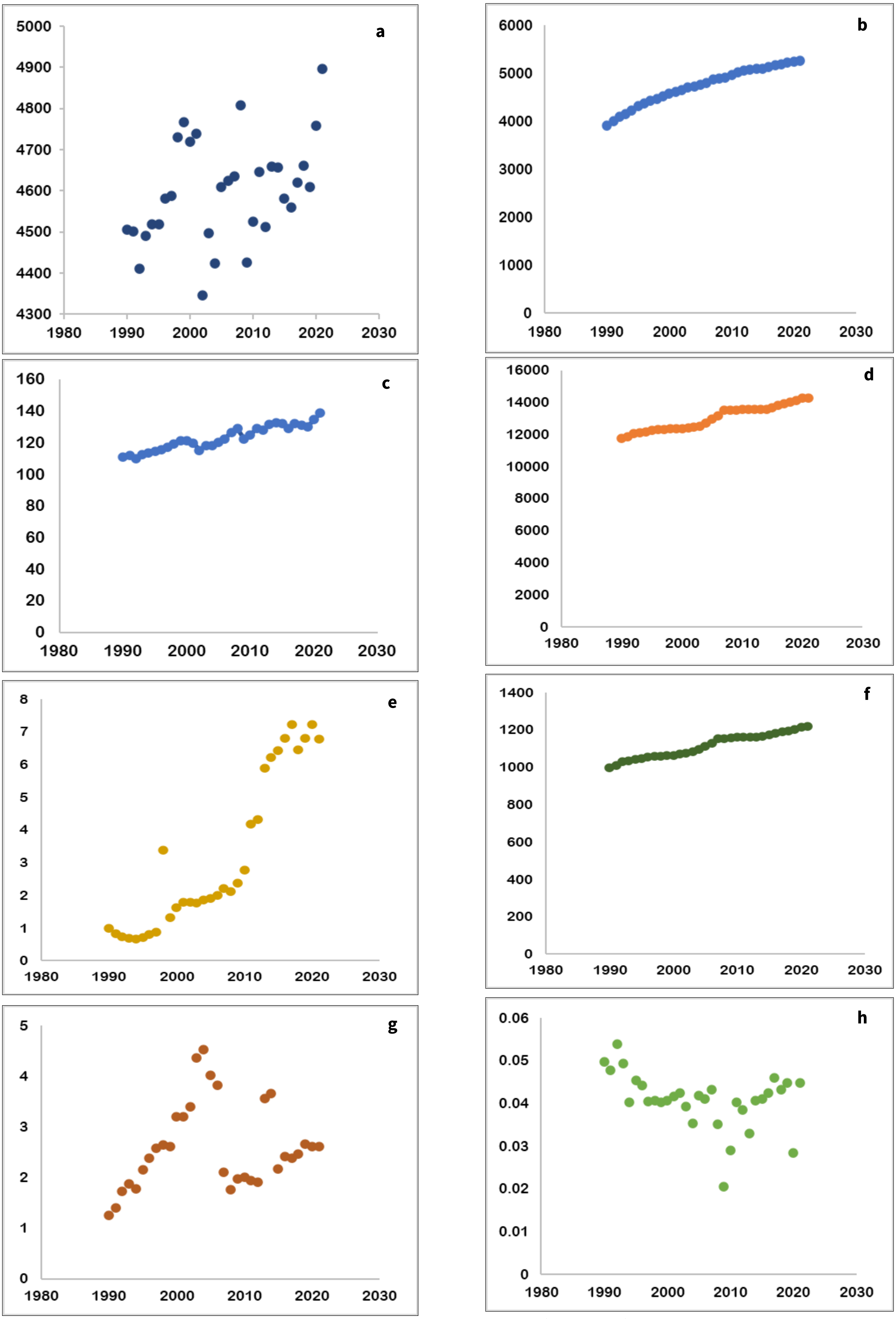

Fig. 5. A concise overview of India's methane emissions from agriculture. The years (1990 - 2021) are represented by the X axis and the methane emission volume (measured in $MtCO_2eq$ per year) is indicated by the Y axis. Fig (a) show the cultivation of rice, the disposal of waste from agri-food systems (b), the burning of crop residues (c), the enteric fermentation process (d), the consumption of food by households (e), the management of manure (f), the use of energy on farms (g) and the manufacturing of pesticides (h).

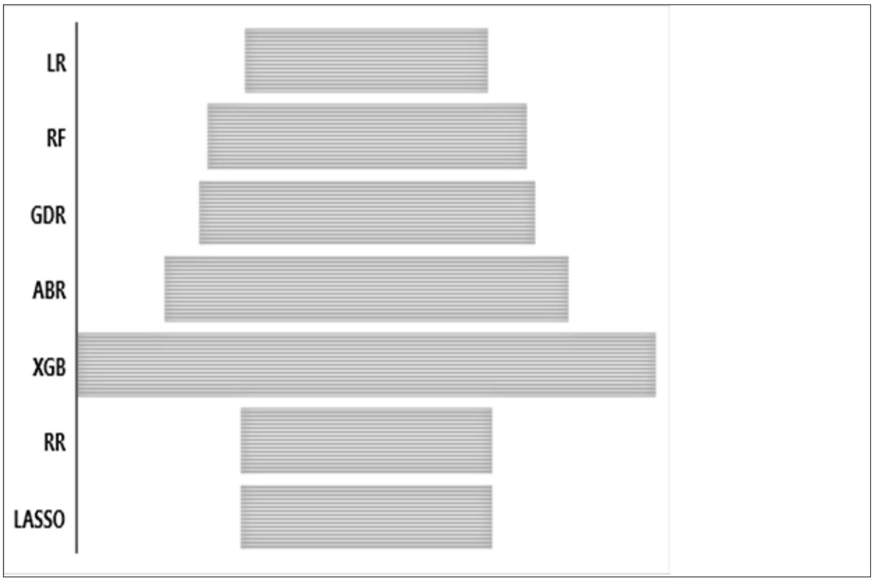

Fig. 6. MAE assessment

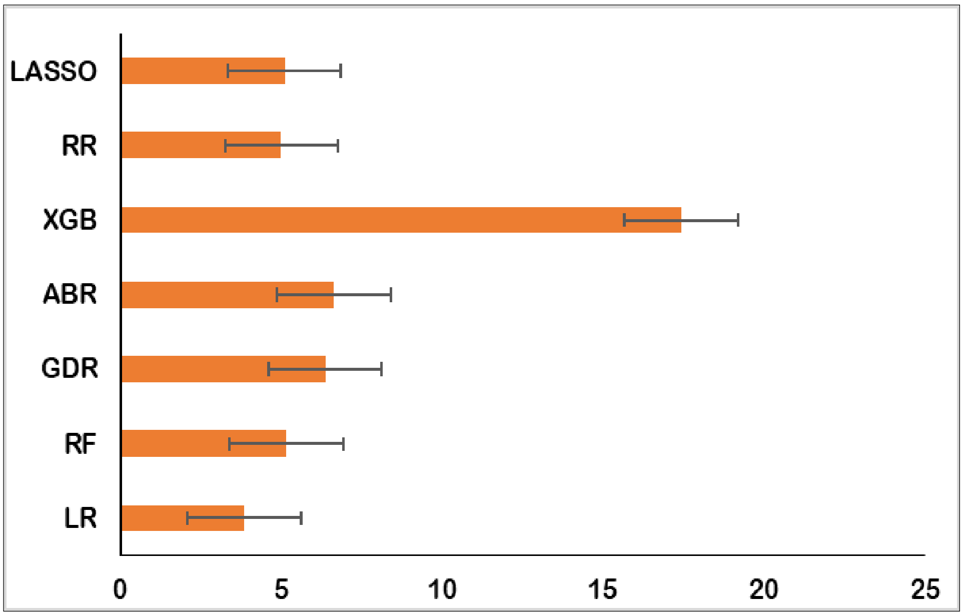

Fig. 7. MSE assessment.

(1.45), along with the highest R^2 (0.98) value, the linear regression model achieved the best forecast accuracy among all models. Its higher training accuracy further validated its superior performance over other models.

This research component focuses on the significant task of projecting methane emissions from India's agricultural activities during the next 10 years, from 2021 - 2030. Using relevant historical data patterns and environmental characteristics, we apply advanced machine learning techniques (LR, RF, XGB, GDR, ABR, RR and LASSO) to predict future emissions. Table 3 and Fig. 10 provide a comprehensive summary of the predicted methane emissions levels, offering insights in potential future trends. To ensure accuracy, we compare these projected values to genuine datasets of methane emissions from agricultural operations between 1990 and 2019. The contrast is shown graphically, with the historical trends displayed beside our model's approximations. Adding the best-performing Linear regression technique aims to increase our forecasts' accuracy and consistency.

Our analysis incorporates dynamic factors and unpredictable settings beyond simple extrapolation. This approach aims to capture the complexity of environmental and agricultural dynamics, providing a deeper understanding of future emissions trajectories. This strategy enables accurate forecasting for the 2021 - 2030 time frame contributing valuable insights to climate change mitigation and policy development (Fig. 11).

Table 2. Evaluation criteria values for all the machine learning models used in this study

Evaluation metrics	MAE	MSE	RMSE	R^2	Test accuracy (%)	Train accuracy (%)
LR	1.45	3.86	1.95	0.988	98	99
RF	1.91	5.15	2.27	0.985	98	99
GDR	2.01	6.36	2.52	0.981	98	99
ABR	2.41	6.64	2.62	0.982	96	99
XGB	3.46	17.43	4.17	0.957	95	99
RR	1.50	5.01	2.23	0.985	98	98
LASSO	1.50	5.10	2.34	0.984	98	98

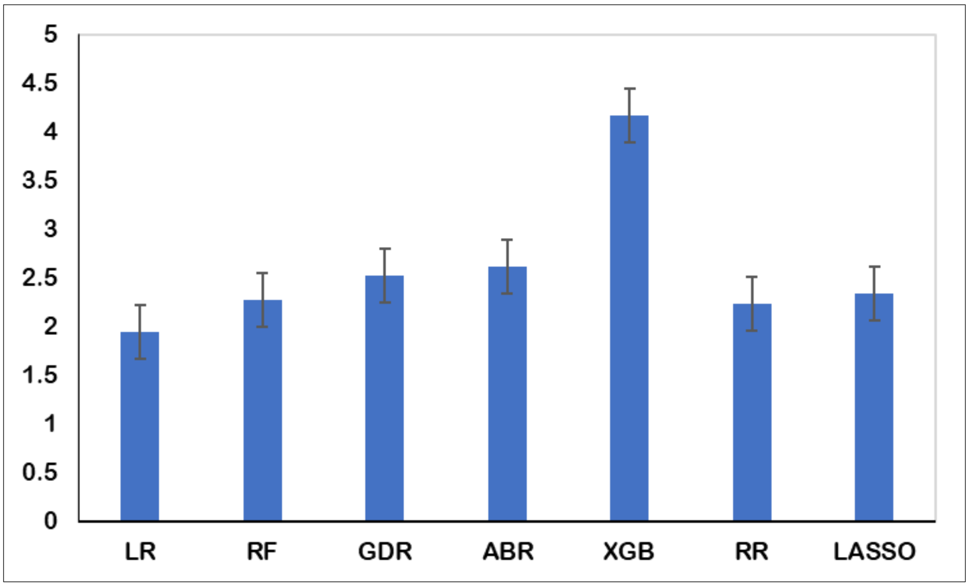

Fig. 8. RMSE assessment.

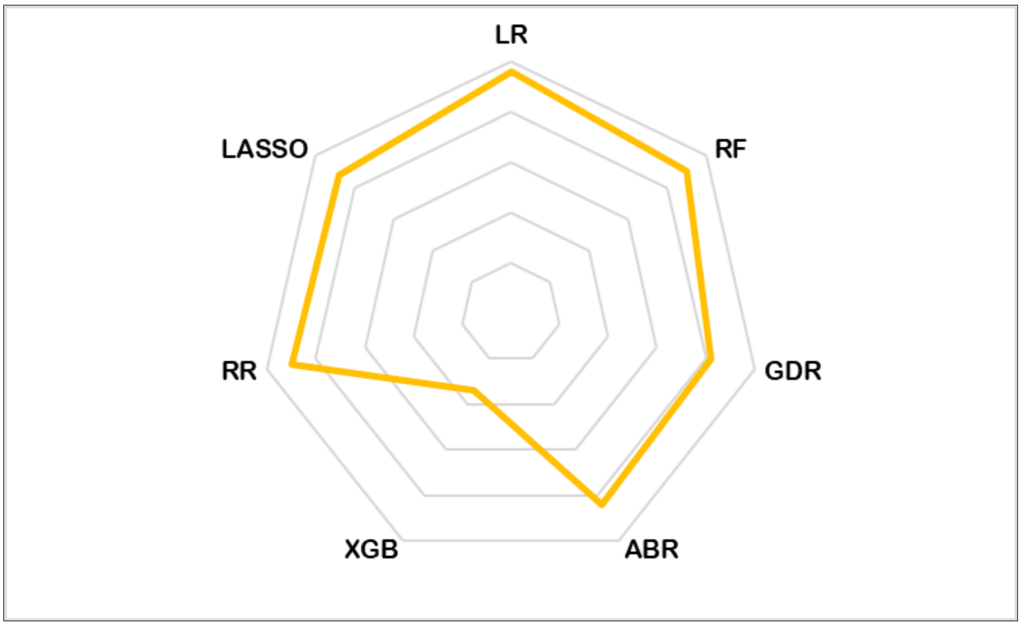

Fig. 9. R^2 assessment.

Table 3. Predicted agricultural methane emissions over the next decade (2021-2030)

Year	Total CH_4 emission, MT
2021	448.134
2022	478.664
2023	468.121
2024	485.859
2025	472.382
2026	487.66
2027	502.188
2028	489.138
2029	492.335
2030	498.894

Conclusion

Our study shows that the linear regression approach demonstrates superior predictive performance in forecasting methane emissions from agricultural operations in India for the period of 2021 - 2030, outperforming other machine learning models. With the highest R^2 (0.98) value and the lowest RMSE (1.95), MSE (3.86) and MAE (1.45) among all models, the linear regression model scored best in terms of forecasting accuracy. This model can be used to forecast methane emissions from agriculture and inform economic strategies aimed at reducing air pollution.

The models developed in this study have demonstrated strong positive predictive performance for forecasting agricultural methane emissions. However, these models are limited to estimating emissions from agricultural fields and do not account for all possible pollution scenarios, such as variations in farming practices or external environmental factors. Although non-linear models like XGBoost and Random Forest typically outperform linear regression in complex datasets, our analysis found that methane emissions from agriculture exhibited a strong linear trend over time. The linear regression model achieved an R^2 of 0.98, outperforming other models due to its simplicity and reduced overfitting risk. However, further research should explore hybrid models that incorporate both linear and non-linear approaches.

Policymakers can use these findings to develop targeted methane reduction strategies in agriculture by promoting low-emission rice cultivation, improved livestock

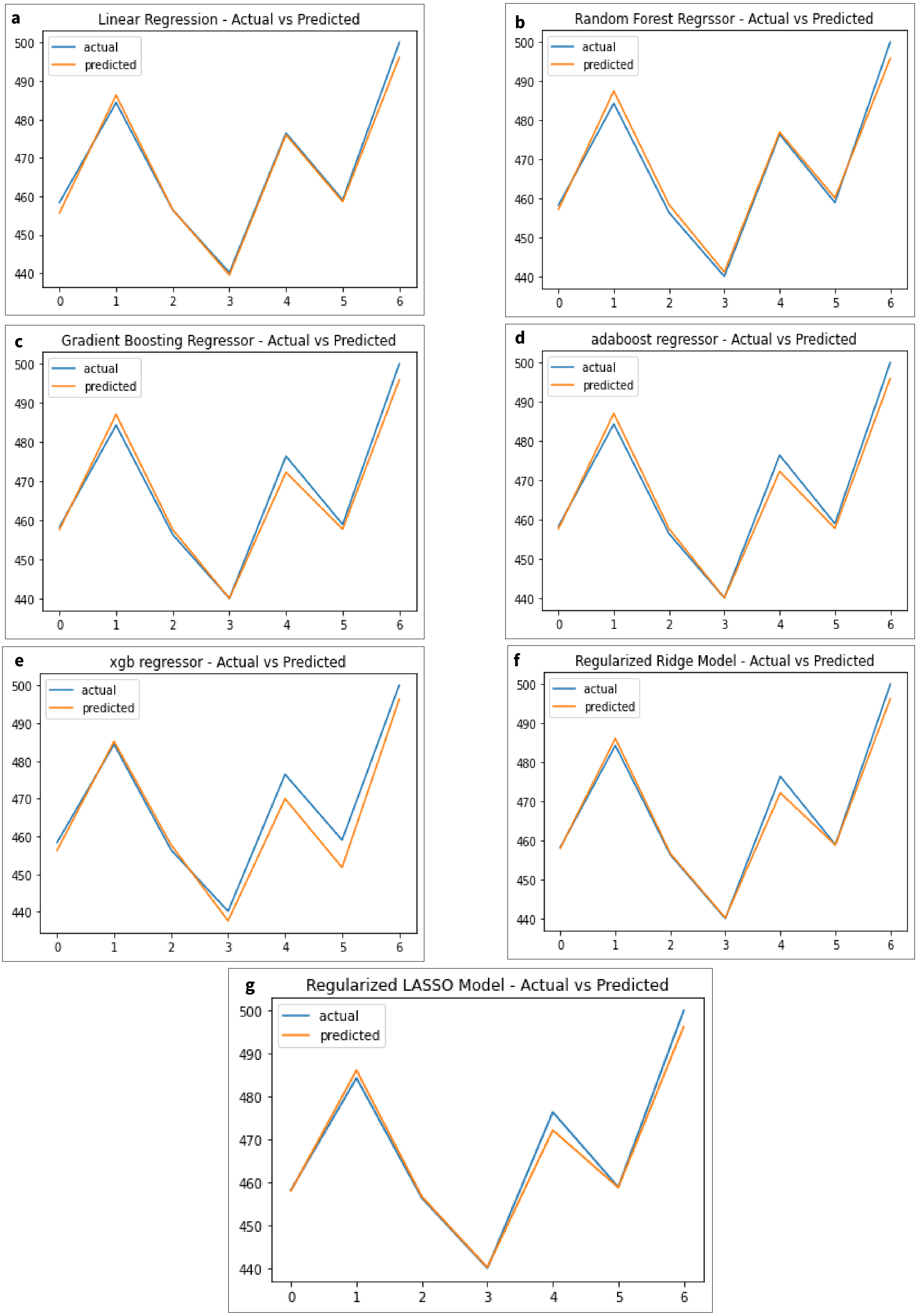

Fig. 10. Line graph of machine learning algorithms used to forecast emissions of methane. The anticipated values are shown on the y-axis in the image, while the actual values are shown on the x-axis. The orange curve shows the outcomes that the model anticipated, whereas the blue curve shows the actual data.

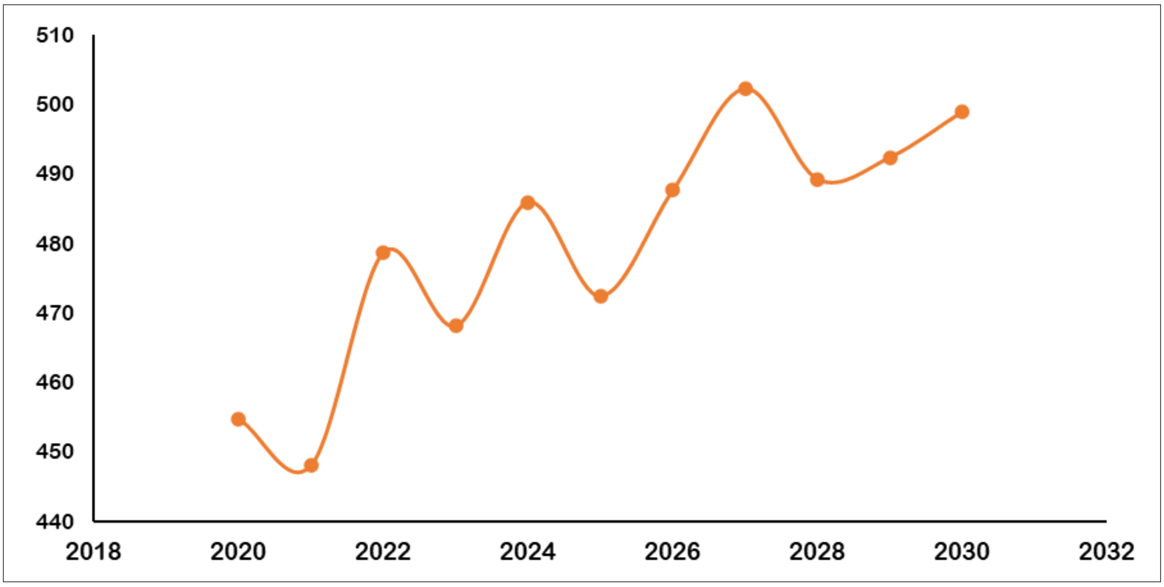

Fig. 11. Predicted methane emission from agricultural activities for the next ten years (2021 - 2030).

feed management and enhanced manure treatment practices. The study's machine learning-based forecasting approach can aid in better monitoring and regulation of agricultural emissions. Integrating these insights into national climate policies can help India align with global methane reduction commitments, improving sustainability and food security while mitigating climate change impacts.

Limitations

This study focuses solely on methane emissions from agricultural activities and does not account for methane from wetlands, waste management, or energy production. Our dataset covers the period 1990 - 2021, but future research should incorporate real-time satellite observations to enhance accuracy. Linear regression provided the best results for our dataset, future studies should explore ensemble learning techniques to model seasonal and regional variations in methane emissions.

Acknowledgements

The authors are thankful to the Tamil Nadu Agricultural University for supporting their research work.

Authors' contributions

VPNB contributed to conceptualization, methodology development, resource acquisition, investigation, data analysis and writing the original draft. KM helped in the statistical analysis, conducting the investigation, formulating the methodology, employing software and writing the original draft. BN assisted with the data analysis, investigation, methodological framework, software application and writing. BV, SA participated in writing, review and editing the manuscript. BV also took part in framing the methodology, software use, writing, review and editing. MR was involved in writing, review and editing. VPNB, MR and DR contributed to writing, review and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interest to declare.

Ethical issues: None

References

1. Sharma SH, Sharma PU. Agricultural production, marketing and food security in India: A peep into progress. Productivity. 2017;58 (2).
2. Chowdhury S, Rubi MA, Bijoy MHI. Application of artificial neural network for predicting agricultural methane and CO2 emissions in Bangladesh. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2021. p. 1–5. https://doi.org/10.1109/ICCCNT51525.2021.9580106
3. Press Information Bureau (PIB). Measures to reduce methane emissions (Internet). Ministry of Environment, Forest and Climate Change; 2023 (cited 2024 Apr 25). Available from: https://pib.gov.in/PressReleaseIframePage
4. Ganesan AL, Schwietzke S, Poulter B, Arnold T, Lan X, Rigby M, et al. Advancing scientific understanding of the global methane budget in support of the Paris Agreement. Glob Biogeochem Cycles. 2019;33(12):1475–512. https://doi.org/10.1029/2018GB006065
5. Global Carbon Project. The global methane budget 2000-2017 (Internet). 2020 Jul 15 (cited 2024 Apr 25). Available from: https://www.globalcarbonproject.org/
6. Dlugokencky E. Trends in atmospheric methane (Internet). NOAA/ESRL; 2020 (cited 2024 Apr 25). Available from: https://www.esrl.noaa.gov/gmd/ccgg/trends
7. Anika OC, Nnabuife SG, Bello A, Okoroafor RE, Kuang B, Villa R. Prospects of low and zero-carbon renewable fuels in 1.5-degree net zero emission actualisation by 2050: A critical review. Carbon Capture Sci Technol. 2022;5:100072. https://doi.org/10.1016/j.ccst.2022.100072
8. Uyar GFU, Terzioglu M, Kayakus M, Tutcu B, Cosgun A, Tonguc G, et al. Estimation of methane gas production in Turkey using machine learning methods. Appl Sci. 2023;13(14):8442. https://doi.org/10.3390/app13148442
9. United Nations Framework Convention on Climate Change (UNFCCC). COP 27: 27th session of the Conference of the Parties (Internet). Government of the Arab Republic of Egypt; 2022 Nov 6-

18 (cited 2024 Apr 20). Available from: https://unfccc.int/event/cop-27

10. FAOSTAT. Food and agriculture data: climate change; agrifood systems emissions, emissions total (Internet). Rome: Food and Agriculture Organization of the United Nations; 2023 (cited 2024 Apr 25). Available from: https://www.fao.org/faostat/en/#home

11. Climate Watch. GHG emissions (Internet). Washington, DC: World Resources Institute; 2024 (cited 2024 Apr 28). Available from: https://www.climatewatchdata.org/ghg-emissions

12. Ghaderzadeh M, Asadi F, Hosseini A, Bashash D, Abolghasemi H, Roshanpour A. Machine learning in detection and classification of leukemia using smear blood images: A systematic review. Sci Program. 2021;2021:9933481. https://doi.org/10.1155/2021/9933481

13. Kim SJ, Bae SJ, Jang MW. Linear regression machine learning algorithms for estimating reference evapotranspiration using limited climate data. Sustain. 2022;14(18):11674. https://doi.org/10.3390/su141811674

14. Mallikarjuna P, Jyothy SA, Sekhar Reddy KC. Daily reference evapotranspiration estimation using linear regression and ANN models. J Inst Eng (India) Ser A. 2012;93:215–21. https://doi.org/10.1007/s40030-013-0030-2

15. Mosre J, Suarez F. Actual evapotranspiration estimates in arid cold regions using machine learning algorithms with in situ and remote sensing data. Water. 2021;13(6):870. https://doi.org/10.3390/w13060870

16. Breiman L. Random Forest. Berkeley: University of California; 2001.

17. Athanasiadis I, Ioannides D. A machine learning approach using random forest and lasso to predict wine quality. Int J Sustain Agric Manag Inform. 2021;7(3):232–51. https://doi.org/10.1504/IJSAMI.2021.118129

18. Vinci A, Zoli L, Sciti D, Melandri C, Guicciardi S. Understanding the mechanical properties of novel UHTCMCs through random forest and regression tree analysis. Mater Des. 2018;145:97–107.https://doi.org/10.1016/j.matdes.2018.02.061

19. Jaiswal JK, Samikannu R. Application of random forest algorithm on feature subset selection and classification and regression. In: 2017 World Congress on Computing and Communication Technologies (WCCCT). IEEE;2017. p. 65–68. https://doi.org/10.1109/WCCCT.2016.25

20. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, NY, USA; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785

21. Upadhyay D, Manero J, Zaman M, Sampalli S. Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids. IEEE Trans Netw Serv Manag. 2020;18 (1):1104–16. https://doi.org/10.1109/TNSM.2020.3032618

22. Osman AIA, Ahmed AN, Chow MF, Huang YF, El-Shafie A. Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J. 2021;12(2):1545–56. https://doi.org/10.1016/j.asej.2020.11.011

23. Xv Y, Sun Y, Zhang Y. Prediction method for high-speed laser cladding coating quality based on random forest and AdaBoost regression analysis. Materials. 2024;17(6):1266. https://doi.org/10.3390/ma17061266

24. Tang J, Henderson A, Gardner P. Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. Analyst. 2021;146(19):5880–91. https://doi.org/10.1039/D0AN02155E

25. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634

26. Aziz S, Chowdhury SA. Analysis of agricultural greenhouse gas emissions using the STIRPAT model: A case study of Bangladesh. Environ Dev Sustain. 2023;25(5):3945–65. https://doi.org/10.1007/s10668-022-02224-7

27. Wang S, Chen Y, Cui Z, Lin L, Zong Y. Diabetes risk analysis based on machine learning LASSO regression model. J Theory Pract Eng Sci. 2024;4(1):58–64.

28. Picard RP, Cook RD. Cross-validation of regression models. J Am Stat Assoc. 1984;79:575–83. https://doi.org/10.1080/01621459.1984.10478083

29. Di Bucchianico A. Coefficient of determination (R^2). In: Encyclopedia of Statistics in Quality and Reliability. Hoboken, NJ, USA: Wiley Online Library; 2008.

30. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7:1247–50. https://doi.org/10.5194/gmd-7-1247-2014

31. Tuchler M, Singer A, Koetter R. Minimum mean squared error equalization using a priori information. IEEE Trans Signal Process. 2002; 50:673–83. https://doi.org/10.1109/78.984761

32. Ritchie H, Rosado P, Roser M. CO_2 and Greenhouse Gas Emissions. OurWorldinData.org (Internet). 2023 (cited 2024 Apr 20). Available from: https://ourworldindata.org/co2-and-greenhouse-gas-emissions

33. Jones MW, Peters GP, Gasser T, Andrew RM, Schwingshackl C, Gutschow J, et al. National contributions to climate change due to historical emissions of carbon dioxide, methane and nitrous oxide since 1850. Sci Data. 2023;10(1):155. https://doi.org/10.1038/s41597-023-02041-1

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.