Chapter 7

Comprehensive analysis of machine learning algorithms for biogas prediction from rice straw

7.1. Background

Anaerobic digestion (AD) comprises a sequence of biochemical and physical processes. It is widely employed for the treatment of diverse organic wastes, such as lignocellulosic biomass, with the dual purpose of producing renewable energy and nutrient-rich digestate [224]. The utilization of AD technology is crucial in the circular bioeconomy and makes a significant contribution towards achieving carbon neutrality [225]. The AD performance depends on multiple operational parameters like temperature, pH, chemical oxygen demand (COD), VFAs, and activity of microbial enzymes to efficiently break down organic substances [226]. Although the AD process has advantages, it frequently encounters instability, which adversely affects biogas output. Considerable endeavours have been undertaken to develop techniques with the objective of managing the AD process, with an emphasis on maintaining process stability and forecasting overall AD performance [227]. Anaerobic process modelling can be a useful technique for predicting crucial process performance factors, such as methane generation. The Anaerobic Digestion Model No. 1 (ADM1) is currently the most robust mechanistic model for predicting biogas generation over time in anaerobic digestion processes. ADM1 model is computationally intensive, making real-time prediction and control not feasible. For full-scale industrial AD processes with different feedstock contents, thorough ADM1 parameter calibration is very difficult [228]. Machine learning (ML)-based models and soft computing techniques have evolved as an alternate method for AD process modelling in order to overcome these limitations [229].

Data analytics can serve as an objective tool to examine the outcomes of experiments and has the potential to modify the nature of the laboratory experiment by building a foundation for analyzing its effectiveness [230]. ML-based AD process modelling is faster than mechanistic models (e.g., ADM1), requires less bio-kinetics, microbiome, and heat/mass transfer information, and does not require model re-calibration on big datasets [231]. Machine learning has the potential to uncover hidden relationships between several input variables and output predictions. Recent studies have demonstrated the application of some ML algorithms to model the non-linear and intricate relationships in anaerobic digestion like Artificial Neural Network (ANN), Random Forest (RF), Linear Regression (LR), Decision Trees, Gradient Boosting etc. [232–235]. ANN and nonlinear regression were used to estimate biogas production using experimental data in a study. High R^2 values (0.9852 for ANN, 0.9878 for regression) confirmed model significance [230]. Holubar et al. (2002) modelled and controlled methane generation from anaerobic continuously stirred tank reactors with varying organic loading rates using ANNs. Results showed that the models could forecast reactor gas generation and composition [236].

This paper aims to develop a complex and intricate relationship between many parameters and biogas/methane production through rigorous and exploratory data analysis of the AD of rice straw data. The objective of this study is to use advanced statistical tools, mathematical processes, and machine learning (ML) to obtain numerical results that demonstrate the effectiveness of the anaerobic digestion (AD) process and the advantages of different physical, chemical, and biological pre-treatments as an environmentally friendly approach to solve the problem of rice straw disposal. De Clercq et al. (2019) created a web-tool to predict biogas output and revenue using regression models (De Clercq et al., 2019a).

7.2. Development of prediction models using Machine Learning algorithms

7.2.1. Data collection

The main aim of this study was to model the primary data of an AD experiment. The experiment for anaerobic digestion of rice straw was conducted for 45 days to evaluate the efficiency of ML algorithms to predict AD performance. The data collected and presented in Chapter 5 has been used here to build prediction models. The independent variables were sample name, day, pH, VFA, activity of xylanase and cellulase enzymes, sCOD, biogas, methane and methane percent and data collection were done every 3 days.

Methane percentage was calculated using the following formula:

$$Percent\ of\ methane\ in\ biogas = \frac{volume\ of\ methane}{volume\ of\ biogas} * 100$$

This gave us a data set consisting of 585 x 10 dimensions (15 days * 13 samples * 3(for triplicates)). For building a regression model the minimum data required is 30 observations for one independent variable and at least 10 observations have to be added for every additional independent variable [238]. The dataset has 7 independent variables and 2 dependent variables (biogas and methane), minimum observations for an acceptable model is 90 (30+(6*10) = 90). This means that 30 observations are needed for the first independent variable and 10 additional observations for each of the next 6 variables. 585 observations will yield a good regression model; the data collection is efficient and will yield potent and practical predictive models. The biogas and methane are the dependent output variables.

7.2.2. Overview of data set

The independent variable ‘Sample’ contains the sample name. The values for ‘Sample’ are “categorical” with type “text” with the following categories: NC for Negative Control,

PC for Positive Control, AU for Autoclaved, PS10, PS20, PS30, PS40 for 10, 20, 30, 40 days PS-pretreated samples respectively, TL10, TL20, TL30, TL40 for 10, 20, 30, 40 days TL-pretreated samples respectively, and NaOH1 and NaOH2 for 1% NaOH and 2% NaOH pretreated samples respectively. The explanation for the sample names is also described in Chapter 5.

The 'Day' variable records the day when the observations were recorded and are of "integer" type. Other integer-type observations are SCOD and biogas. Methane volume was calculated by performing gas chromatography of the biogas sample collected and is a "floating-point" number. All other variables (pH, VFA, xylanase and cellulase) are "floating-point" numerical features. The data is carefully and consistently recorded; hence, there are no NULL or NA-type entries present. Hence, null handling is not required to perform the subsequent study.

7.2.3. Data transformation

The categorical values of the sample name in the column 'Sample' were converted into numerical variables by adding Boolean parameters with the value 0 when a sample is not a constituent of the category and 1 when the observed sample belongs to the category, e.g., if the sample belonged to NC category the value of 'isNC' parameter was 1 and for all other newly introduced parameters were 0 or if the sample belonged to NaOH2 category, then value for 'isNaOH2' parameter was 1 and for all other introduced parameters were 0. This method was used to convert categorical variables to numerical variables because, as per the experimental hypothesis, the pretreatment methodology should impact the biogas generation, and on further analysis, it was concluded to be correct, as mentioned in *Chapter 5*. In addition to the categorical to numerical conversion, the categorical variable was removed from the dataset as the impact is captured in the

newly introduced numerical variables, and regression models do not work on categorical values.

The values were further normalized using the following formula:

$$X_{normalised} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where,

$X_{normalised}$ gives the normalized value of the variables, X represents the original value of the variable, X_{min} is the minimum value, and X_{max} is the maximum value of the variable in the dataset

Normalization is a method employed to rescale and standardize the range of independent variables or features within a dataset [239]. Due to significantly high magnitudes of values observed in some parameters (like sCOD), most of the variability was explained by those parameters. Once the data is normalized, values of all the predictors become comparable, and a more balanced and sophisticated model is achieved, giving equal consideration to all the predictors.

7.2.4. Data visualisation

Scatter plots are a good visual tool to identify any visually discernible patterns in the data. Using the GGPLOT library in R, pair-wise scatter plot was generated that showed the inter-relationship between Day, pH, xylanase, cellulase, SCOD, methane percent, methane and biogas. The aggregate view of pair-wise plots is a good method to showcase the relationship between different controlling variables with each other and with biogas as well as methane.

7.2.5. Data correlation analysis

For statistical analysis of variables, Pearson's coefficient has been employed. Pearson's coefficient, often denoted as "r," has been deemed as a milestone for quantifying the extent of the linear relationship between two variables. Attributing to its interpretability and simple application, the measure is a go-to tool for data scientists. The value of *r* ranges from -1 to +1, thus providing a unitless statistic to signify strength of correlation. The absolute magnitude of *r* denotes the strength of the correlation, and the symbol signifies the direction of the association, i.e., a negative r means that there exists a negative linear interdependence between variables of the study and that if one increases, the other decreases. The formula for Pearson's correlation coefficient between variables X and Y based on a set of data pairs (x_i, y_i) is provided by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

Where,

n is the total number of data pairs, x_i and y_i represent the values of individual data of the variables X and Y, $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y variables, respectively

The correlation coefficient is determined by normalizing the covariance of the variables with the product of their standard deviations.

Sample names are non-numerical categorical factors that would significantly impact biogas and methane production as per our experiment hypothesis. The chi-square test is used to calculate the correlation between categorical variables. Since biogas and methane are numerical variables, converting them into categorical variables is required. This can

be done by dividing the total values into 3 parts and giving them category names as ‘Low,’ ‘Medium’, and ‘High.’ The ranges are defined as

Low = *0* to *(max value – min value)/3*

Medium = *(max value – min value)/3* to *2*(max value – min value)/3*

High = *2*(max value – min value)/3* to *max value*

7.2.6. Machine learning algorithms

Checking auto-correlation is needed before machine learning algorithms are trained. It is expected that methane would be highly correlated to biogas. Since both biogas as well as methane are dependent variables and would highly impact the training of ML models, thus reducing the significance of independent variables. Variance inflation factor (VIF) is used to test the assumption and detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis. Multicollinearity inflates the variance and error. It makes the coefficient of a variable consistent but unreliable. VIF yielded that methane, methane percentage, and biogas together resulted in high autocorrelation. Hence, the data set was modified into two datasets, one with biogas as the outcome variable and devoid of methane and the other with methane as the independent variable. Methane percent was removed from consideration since it is a calculated field and not directly used as a variable of interest. The two datasets were divided into a 75:25 ratio to train the ML models, with 75% data as the training set and 25% data to test/validate the accuracy of the model. The actual values of methane and biogas production in mL/g VS for the different pretreatments can be referred from Figure *5.1* and Figure *5.2*.

7.2.6.1. Linear regression (LR)

Linear regression is an ML approach to forecast a continuous result variable by considering one or more predictor variables. Within the realm of ML, the linear regression model undergoes training using a dataset in order to acquire knowledge about the connection between the input features and the target variable. The model postulates a linear correlation between the characteristics and the objective variable [240]. The equation for the linear regression model is as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where,

Y is the target variable, X is the predictor variable, β_0 represents the intercept with β_1 giving the slope and ε shows the error.

7.2.6.2. Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised ML technique that is typically used for classification and regression applications. SVM works by identifying the best hyperplane that divides data points belonging to various classes. In the instance of classification, the goal is to design a hyperplane that maximizes the margin between various classes [234]. Hyperplanes divide data into two classifications in two dimensions. It becomes a hyperplane in higher dimensions. The margin is the hyperplane's distance from each class's nearest data point. The hyperplane with the greatest margin is SVM's goal. Support vectors, the data points nearest to the hyperplane, define the margin. These are crucial in choosing the best hyperplane. SVM maximizes margin and minimizes classification error. Quadratic programming is used to discover the best hyperplane coefficients [241].

7.2.6.3. Decision trees (DT)

Common machine learning algorithms for classification and regression are decision trees. They recursively partition data by features and make judgments at each tree node [231]. The decision-making process is a flowchart with internal nodes representing feature-based decisions, branches representing outcomes, and leaf nodes representing final decisions or predictions. The tree's top node, representing the complete dataset. Feature selection divides it into child nodes. Nodes representing feature-based judgments. Each internal node has branches for feature values. Edges linking nodes. The parent node's decision determines each branch. Bottom nodes of the tree. In classification and regression, leaf nodes represent class labels or numeric values. Each internal node decides depending on a feature and threshold value. Decision criteria maximize data purity or homogeneity in each partition. Entropy and information gain help Decision Trees choose data splitting characteristics. Information gain measures how well a feature reduces uncertainty [242].

7.2.6.4. Random forest (RF)

Random Forest improves prediction by combining many decision trees. Classification and regression are its goals. The approach trains a "forest" of decision trees and outputs their average regression prediction or mode classification prediction. Bootstrapping the training dataset creates numerous decision trees in Random Forest. Each tree is trained on a distinct data subset [243]. The decision tree splits a random subset of features at each node. This diversifies the trees, preventing strong correlation. Classification uses "votes" from each tree, and the class with the most votes is predicted. Average tree predictions are used in regression. Random Forest bags predictions from various models to reduce overfitting and increase generalization. Multiple decision trees in Random Forest

typically yields great accuracy. Ensembles reduce overfitting compared to individual decision trees. Random Forest handles missing data. It can aid feature selection by revealing feature relevance [244].

Long et al. (2021) used various ML methods to forecast methane yield from AD genomic data and operational parameters. RF predicted the best results. RF also offered vital information regarding the most relevant phyla to improve process control techniques [245].

7.2.6.5. Gradient boosting (GB)

Gradient Boosting sequentially combines the predictions of numerous weak models, usually decision trees, to create a powerful predictive model. Gradient Boosting iteratively fits new models to earlier errors to improve model weaknesses. It is popular for regression and classification. Weak learners like shallow decision trees or simple linear models are utilized in gradient boosting. These "weak" models perform marginally better than random guessing. Gradient Boosting creates trees successively, correcting errors from the pooled forecasts of the previous trees. Every new tree is trained on the combined model's residuals (the disparities between true values and forecasts). This method is directed by an error-quantifying loss function. The learning rate determines how much each new tree affects the model. Lower learning rates require more trees but improve generalization. The loss function is minimized by gradient descent. Calculates the loss gradient relative to the model's anticipated values and modifies predictions in the opposite direction. eXtreme gradient boosting (XGBoost) was also tested for biogas and methane predictions.

A study analyzed 386 anaerobic digestion facilities in Germany and the USA, using data envelopment analysis and a stochastic gradient boosting classification model. Efficiency

scores varied (0.82 in Germany, 0.46 in the USA). Key determinants included pre-treatment stages/types, digester technology, and co-digestion presence/absence [246].

7.2.7. R-squared to evaluate the model performance

To test the performance of the ML models, R squared (R^2) value was used and the formulae is mentioned below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i^{exp} - Y_i^{pre})^2}{\sum_{i=1}^{n}(Y_i^{exp} - \bar{Y})^2}$$

where n is the number of samples, Y_i^{exp} is the true value of the observation, Y_i^{pre} is the value predicted by the model, and $\bar{Y}$ is the mean of the sample.

All the models were applied using RStudio.

7.3. Results and discussion

7.3.1. Data visualization

The categorical values of the sample name in the 'Sample' column were transformed into numerical variables by introducing Boolean parameters. These parameters have a value of 0 when a sample does not belong to a particular category, and a value of 1 when the observed sample is part of the category as shown in *Figure 7.1*

Samples	isNC	isPC	isAU	isPS1	isTL0	isPS20	isTL20	isPS30	isTL30	isPS40	isTL40	isNaOH1	isNaOH2
NC	**1**	0	0	0	0	0	0	0	0	0	0	0	0
PC	0	**1**	0	0	0	0	0	0	0	0	0	0	0
AU	0	0	**1**	0	0	0	0	0	0	0	0	0	0
PS10	0	0	0	**1**	0	0	0	0	0	0	0	0	0
TL10	0	0	0	0	**1**	0	0	0	0	0	0	0	0
PS20	0	0	0	0	0	**1**	0	0	0	0	0	0	0
TL20	0	0	0	0	0	0	**1**	0	0	0	0	0	0
PS30	0	0	0	0	0	0	0	**1**	0	0	0	0	0
TL30	0	0	0	0	0	0	0	0	**1**	0	0	0	0
PS40	0	0	0	0	0	0	0	0	0	**1**	0	0	0
TL40	0	0	0	0	0	0	0	0	0	0	**1**	0	0
NaOH1	0	0	0	0	0	0	0	0	0	0	0	**1**	0
NaOH2	0	0	0	0	0	0	0	0	0	0	0	0	**1**

Figure 7.1. Boolean conversion of categorical variables to numerical variables

Figure 7.2 presents the pairwise scatter plot to identify any visually discernible patterns in the data. It shows the inter-relationship between day, pH, xylanase, cellulase, sCOD, methane percent, methane and biogas. Pair-wise graphs show how controlling factors interact with one other, biogas, and methane in the aggregate view. A very apparent trend is a positive slope between biogas and methane which is understandable that higher biogas will result in better methane yield as well. Another clear trend is between day and methane and day and biogas where the later days of the experiment show diminishing biogas and methane production. The experiment maintains near neutral pH with slightly acidic values yielding better biogas production. Another noteworthy observation is that pH value increases with passing days i.e. the media becomes more basic as the time progresses. For lower values of xylanase and cellulase activity, the biogas production varies from low to high but for higher values of the enzymes the biogas generated is relatively lower as observed by a left skewed distribution. A similar pattern is observed with methane with xylanase and methane with cellulase activity. This may be attributed to the fact that xylanase and cellulase enzyme breakdown hemi-cellulose and cellulose respectively.

Methanogens work on simple carbohydrates and break-down of complex carbohydrates to simpler carbohydrates is the focus of microbial community when the value of the two enzymes is high.

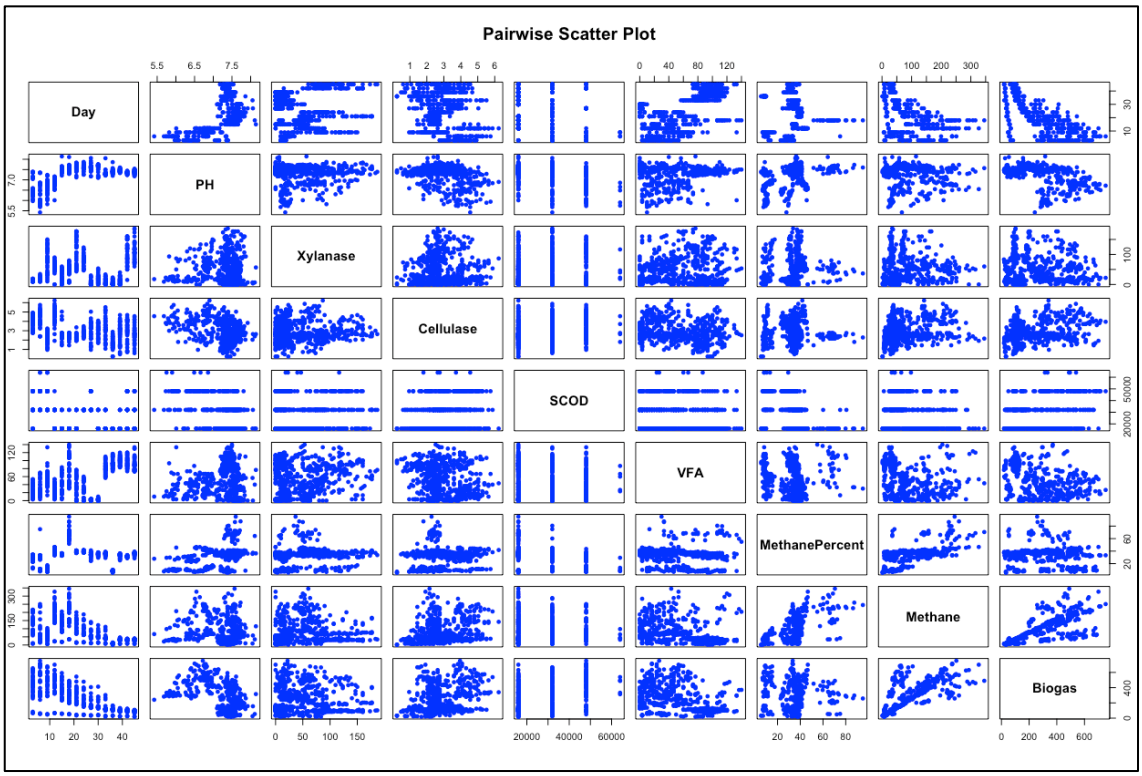

Figure 7.2. Pairwise scatter plot for data visualisation showing inter-relationships between different parameters

7.3.2. Data correlation analysis using a heat map

In the heatmap shown in *Figure 7.3*, the green color shows a positive correlation, and the red color shows a negative correlation. The darkness of color shows the magnitude of Pearson's coefficient. pH and day are positively correlated which means that pH increases as the days proceed in AD. VFA and day are also positively correlated means as the days proceed, VFA content increases. In contrast, biogas and methane have a strong negative correlation with the day, which means that with an increase in day, the biogas as well as methane production diminishes. sCOD and cellulase enzyme activity have a negative

correlation with day while xylanase has a positive correlation with day. Biogas and methane have a strong negative correlation with isNC, which is explicable from the fact that negative control was expected to produce less biogas and methane because of the absence of feedstock. Any significant correlation (>10%) between the sample and biogas or sample and methane is for samples pretreated with NaOH and is mildly positive in both cases. Methane percent is highly correlated with pH. Cellulase and sCOD are positively correlated with Biogas and VFA is negatively correlated. Values have been listed on the heatmap to aid in the interpretation.

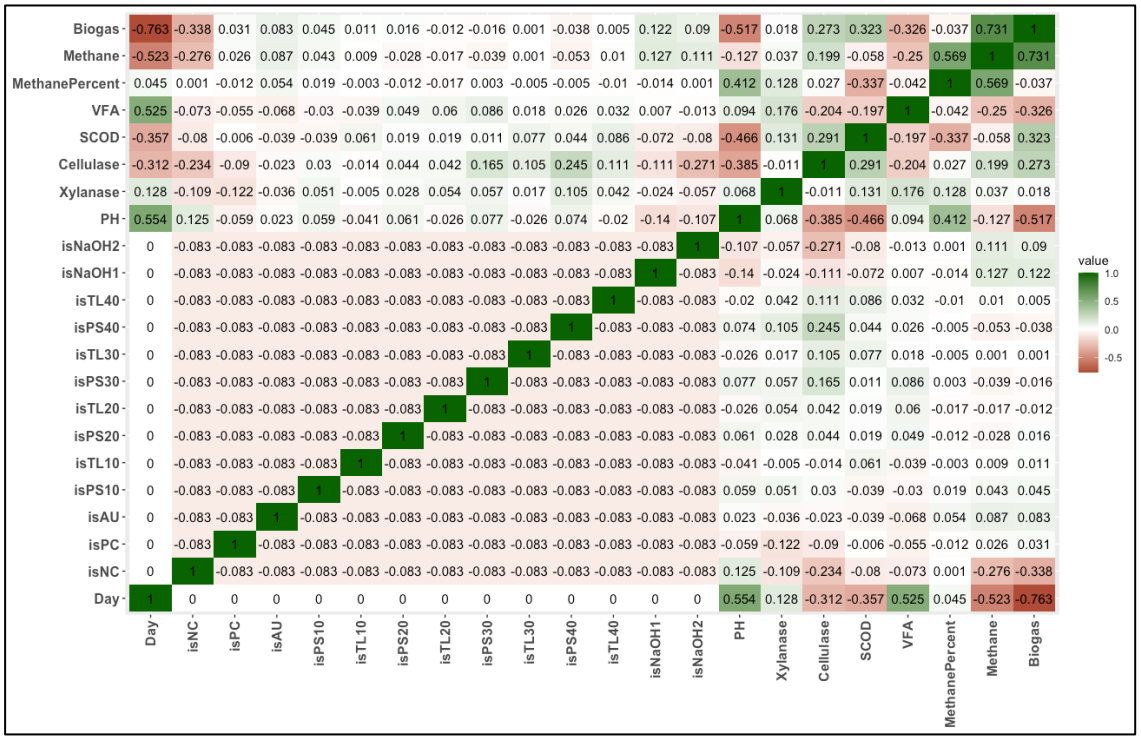

Figure 7.3. The correlation matrix heatmap displays the Pearson correlation coefficient values for all analysed parameters, with positive values shown in green and negative ones in red. The range of values is from -1 to 1. A value of -1 represents a perfect negative linear relationship between variables, a value of 1 indicates a perfect positive linear relationship between variables, and a value of 0 indicates no link between the variables being analysed

A chi-square test was also conducted, and the p-value of the chi-square test is 3.535e-12, which is significantly lower than 0.05; hence, it shows a strong correlation. Thus, it can be concluded that the pre-treatment method significantly impacted biogas production. The above transformation was performed for methane values as well, and the chi-square test revealed that the p-value was 0.004, which in turn shows that a significant correlation is present between the pre-treatment method and methane production.

7.3.3. Comparison of ML models

ML models were applied to predict the production of both biogas and methane. The results of actual and predicted values of biogas and methane were analysed by linear regression, support vector machine, decision trees, random forest, gradient boosting, and eXtreme gradient boosting as shown in *Figure 7.4* and *Figure 7.5*. The regression models were utilized to evaluate the efficacy of the algorithms. The results show that XGBoost performed the best in predicting both biogas and methane with 92% and 91% accuracy. XGBoost is an ensemble learning strategy that combines predictions from numerous weak models to get more accurate forecasts [247]. The technique is highly praised and frequently used in machine learning due to its effectiveness in handling extensive datasets and its capacity to deliver cutting-edge results, especially in tasks like regression analysis. The accuracy from the different ML models used for biogas prediction was in the order: XGBoost (92%) > GB (91%) = RF (91%) > SVM (85%) > DT (79%) > LR (76%). The accuracy followed the order XGBoost (91%) > GB (83%) > RF (87%) > SVM (63%) > DT (615) > LR (56%) for methane prediction. It was observed that the prediction models predicted smaller values of biogas and methane more accurately as compared to larger values.

Biogas production is an ecologically sound and sustainable approach to produce energy from biological waste. Precise forecasting of biogas and methane generation is essential for optimizing production procedures, improving efficiency, and reducing environmental consequences. This study aimed to assess the efficacy of several machine learning algorithms in forecasting these production values.

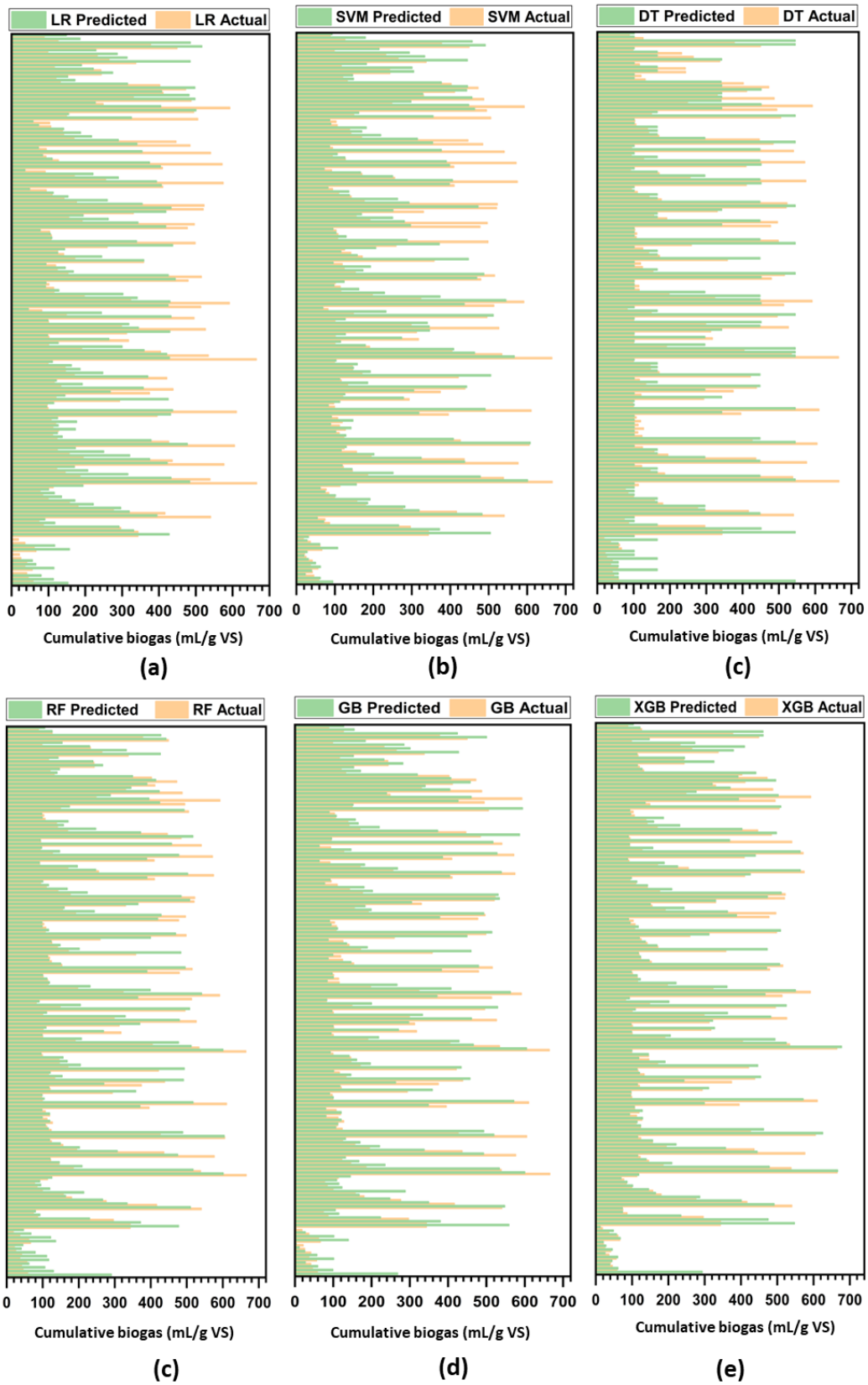

Figure 7.4. Comparison between the actual and predicted biogas production values from ML models: (a) LR, (b) SVM, (c) DT, (d) RF, (e) GB, and (f) XGB

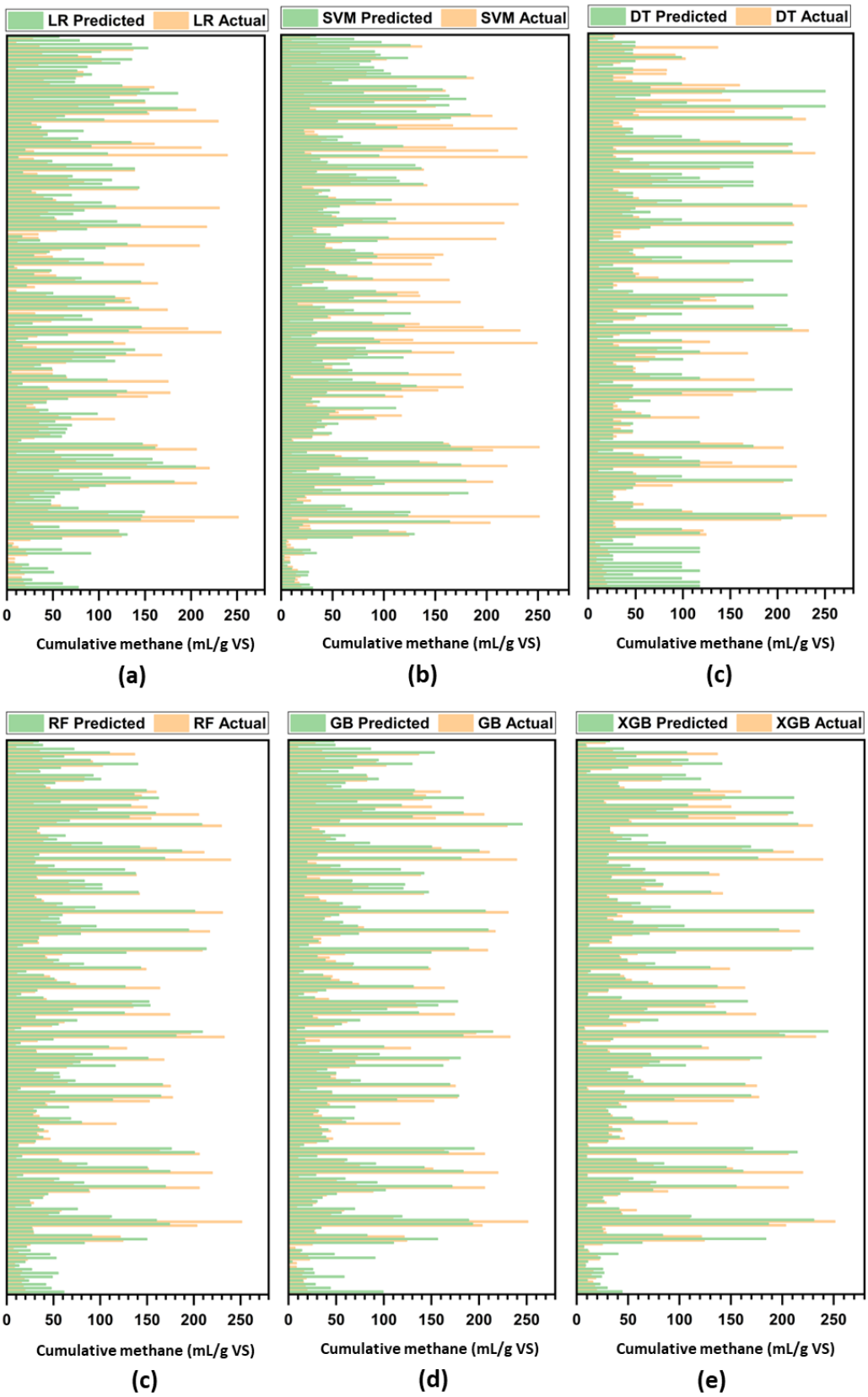

Figure 7.5. Comparison between the actual and predicted methane production values from ML models: (a) LR, (b) SVM, (c) DT, (d) RF, (e) GB, and (f) XGB

XGBoost's effectiveness in predicting biogas and methane production can be due to its capacity to handle non-linear connections, conduct feature importance analysis, and employ regularization techniques. The results of performance accuracy displaying R^2 and correlation between actual and predicted values of biogas and methane are shown in *Table 7.1.*

Table 7.1. Comparison of different ML models to predict the production of biogas and methane (R^2 values range from 0 to 1, where 1 indicates a perfect fit), cor is for testing correlation between actual and predicted values, a value closer to 1 represents a good match

ML algorithm name	**Biogas**		**Methane**	
	R-squared (R^2)	**cor* (actual, predicted)**	**R-squared (R^2)**	**cor* (actual, predicted)**
Linear Regression (LR)	0.76	0.87	0.56	0.75
Support Vector Machines (SVM)	0.85	0.92	0.63	0.8
Decision Trees (DT)	0.79	0.89	0.61	0.8
Random Forest (RF)	0.91	0.96	0.87	0.94
Gradient Boosting (GB)	0.91	0.95	0.83	0.91
eXtreme Gradient Boosting (XGB)	0.92	0.96	0.91	0.95

*cor refers to the correlation between actual and predicted values

7.4. Conclusion

This study used advanced machine learning (ML) to obtain numerical results that demonstrate the efficacy of anaerobic digestion (AD) and the benefits of physical, chemical, and biological pre-treatments as an environmentally friendly rice straw disposal solution. The chi-square test gave $p < 0.05$ showing a strong correlation between categorical variables. eXtreme Gradient Boosting performed the best in predicting biogas and methane values. The ML models employed for biogas prediction achieved the following accuracy rates: XGBoost (92%), GB (91%) and RF (91%) performed equally well, SVM achieved 85%, DT 79%, and LR 76% accuracy. The methane prediction accuracy ranked as follows: XGBoost (91%) > Gradient Boosting (83%) > Random Forest (87%) > Support Vector Machine (63%) > Decision Tree (61.5%) > Logistic Regression (56%). The observation revealed that the prediction models demonstrated greater accuracy in predicting smaller values of biogas and methane as opposed to higher values.