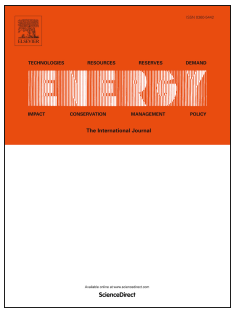
Journal Pre-proof



Predicting the Methane Production of Microwave-Pretreated Anaerobic Digestion of Food Waste: A Machine Learning Approach

Rohit Gupta, Cameron Murray, William T. Sloan, Siming You

PII: S0360-5442(25)02255-8

DOI: https://doi.org/10.1016/j.energy.2025.136613

Reference: EGY 136613

To appear in: *Energy*

Received Date: 20 July 2024

Revised Date: 27 April 2025

Accepted Date: 15 May 2025

Please cite this article as: Gupta R, Murray C, Sloan WT, You S, Predicting the Methane Production of Microwave-Pretreated Anaerobic Digestion of Food Waste: A Machine Learning Approach, *Energy*, https://doi.org/10.1016/j.energy.2025.136613.

Predicting the Methane Production of Microwave-Pretreated Anaerobic Digestion of Food Waste: A Machine Learning Approach

Rohit Gupta[1, #], Cameron Murray[2, #], William T. Sloan[2], Siming You[2, *]

[1] UCL Mechanical Engineering, University College London, London WC1E 7JE, UK

[2] James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK

[#] The authors contribute equally

*Corresponding author details: siming.you@glasgow.ac.uk

Submitted to ***Energy***

August 2024

Abstract

Anaerobic digestion (AD) is a widely adopted waste management strategy that transforms organic waste into biogas, addressing both energy and environmental challenges. Feedstock pretreatment is crucial for enhancing organic matter breakdown and improving biogas yield. Among various techniques, microwave (MW) irradiation-based pretreatment has shown significant promise. However, the optimization of MW-assisted AD processes remains underexplored, necessitating predictive tools for process simulation. Machine Learning (ML) has recently emerged as a powerful alternative for predicting and optimizing AD performance. In this study, an ML-driven pipeline was developed to predict methane yield based on food waste (FW) composition, AD reactor parameters, and MW pretreatment conditions. A range of data preprocessing techniques and ML models (linear, non-linear, and ensemble) were systematically evaluated, with model performance assessed via hyperparameter-optimized cross-validation. The most accurate models (non-linear and ensemble) achieved $R^2 > 0.91$ and RMSE < 35 mL/g volatile solids (gVS), whereas linear models underperformed ($R^2 < 0.71$, RMSE > 70 mL/gVS). Support Vector Machine (SVM) emerged as the best-performing model, with R^2 ~0.94 and RMSE ~34 mL/gVS. Beyond predictive accuracy, this study offers novel insights into MW pretreatment's role in AD efficiency. Permutation feature importance (PFI) analysis revealed that while MW pretreatment enhances methane yield, its effects are secondary to reactor pH and FW composition. This suggests that MW treatment primarily facilitates substrate disintegration but does not drastically alter biochemical methane potential unless coupled with optimized reactor conditions. Additionally, minor fluctuations in MW pretreatment time and temperature were found to have negligible impacts on methane production, indicating a level of operational flexibility in MW-based AD processes. These findings provide a refined understanding of MW pretreatment's practical implications, guiding process design for improved scalability and industrial application.

Keywords: Anaerobic Digestion; Microwave Pretreatment; Food Waste; Machine Learning; Process Model

1 Introduction

With the continued urbanization across the globe, municipal waste production is expected to increase by 70%, resulting in 3.4 billion metric tons by 2050, adding significant pressure on waste management [1]. The organic fraction of municipal waste (OFMSW) typically comprises food waste (FW). As per the UN Food and Agriculture Organization (FAO), 1.3 billion tonne of FW is globally generated each year, typically disposed of via incineration, landfilling, and compositing [2]. This exacerbates the direct greenhouse gas emissions associated with FW disposal, jeopardizing the UN SDG 13 (*i.e.*, climate action). Biogas and digestate production via Anaerobic Digestion (AD) of FW has improved waste valorisation while facilitating a circular economy.

AD uses microbial communities to decompose organic and moisture content-rich FW substrates to produce biogas containing 55-70% methane, a promising source of clean energy production [3]. The semi-solid by-product, digestate is rich in nitrogen and phosphorus-based nutrients, which serve as a potential biofertilizer. AD is a multi-step complex bio-kinetic process, consisting of four sequential stages: hydrolysis, acidogenesis, acetogenesis, and methanogenesis [4]. The methane yield from an AD process is affected by feedstock compositions, bioreactor operating conditions, reactor design, inoculum type, etc, optimization of which is a challenging task. Hydrolysis is one of the slowest stages and determines the organic matter decomposition, ultimately regulating methane yield [5].

Feedstock pretreatment accelerates hydrolysis, enhancing substrate solubilization, biodegradability, and expediting organic waste decomposition. Traditional pretreatment methods encompass: (a) chemical (*e.g.*, saponification and alkali treatments), (b) mechanical (*e.g.*, ultrasonic, extrusion, and grinding), (c) thermal (*e.g.*, steam explosion and hydrothermal processes), or (d) biological (*e.g.*, compositing, fungal, and enzymatic methods) [6]. Microwave (MW)-assisted pretreatment has emerged as a promising thermal method for enhancing anaerobic digestion (AD) processes. MW-assisted pre-treatment offers advantages such as rapid heating rates, improved energy efficiency, and uniform heating [7]. This technique facilitates the release of organic matter from complex substrates like food waste (FW) into the soluble phase, increasing the biodegradable fraction available to microorganisms. MW pretreatment operates at powers ranging from 440 to 500 W, temperatures between 30°C and 160°C, and durations of 1 to 10 minutes [8].

However, MW pretreatment presents specific challenges that require careful consideration. Excessive temperatures (above 160°C) or prolonged treatment times can induce

the Maillard reaction, producing recalcitrant compounds that inhibit microbial activity, thereby reducing AD efficiency and biogas production [9]. Additionally, the non-thermal effects of microwaves and their mechanisms remain subjects of ongoing research and debate. A comprehensive understanding of these effects is crucial for optimizing MW pretreatment conditions. Furthermore, the energy consumption associated with MW pretreatment is a critical factor; the energy input must not outweigh the benefits gained in biogas production. Therefore, it is imperative to optimize MW pretreatment parameters—such as power, temperature, and duration—while considering their holistic impact on methane production and overall process efficiency [10]. Addressing these MW-specific challenges is essential for the effective integration of MW pretreatment in AD systems.

FW is characterized by high moisture and organic content, making it an ideal substrate for the MW-AD process. Nevertheless, the geographical variability of food habits makes FW a complex AD feedstock. This affects their digestibility, hydrolysis rate, and decomposition time, ultimately varying the methane production [11]. MW-based precise uniform heating facilitates enzymatic reaction for breaking complex organic matters, maximizing the biogas yield of AD. For example, varying the pretreatment temperature across a range of 70, 120, and 150°C improves the biogas yield by 2.7%, 24%, and 11.7% respectively [12]. Nevertheless, increasing the temperature beyond a threshold slows down the decomposition rate due to the formation of complex polymers (*e.g.*, melanoidins), which impart an inhibitory effect on the AD reactor. Other investigations have indicated the importance of optimizing MW time and temperature, simultaneously [11]. Although a slower heating rate (HR, 1.9 °C/min) resulted in faster digestibility (due to gradually cell decomposition and lower chances of inhibitory compounds formation from thermal shock), the anaerobic biodegradability improved at a faster HR (7.8°C/min). MW pretreatment at HRs 1.9 and 3.9 °C/min increased the biogas production by 14-fold for the soluble fraction. In contrast, for the whole fraction of FW, HR = 7.8 °C/min improved the biogas yield, suggesting the necessity of transient MW time control for MW-AD [11].

In parallel to the pretreatment parameters, other routinely controlled AD process attributes are temperature, pH, scale of operation (i.e., reactor volume), hydraulic retention time (HRT), *etc*. Meanwhile, feedstock properties such as total solid (TS), volatile solid (VS), and carbohydrate (%C), protein (%P), and lipid (%L) contents are essential components that regulate methane production [13]. To improve the process efficiencies and understand the whole-system operation of the AD process a range of mathematical models have been developed, among which the Anaerobic Digestion Model 1 (ADM1) is one of the most

sophisticated biokinetic models [14]. Nevertheless, the intricate nature of the model limits its applicability to real-time AD reactor control systems, moreover, the ADM1 requires extensive model calibration before industrial implementation [15]. To circumvent the drawbacks of ADM1, machine learning-based methane yield prediction models have rapidly emerged over the past few years [16].

Frequent choices for ML models have been Artificial neural network (ANN), K-nearest neighbour (KNN), Linear regression (LR), ElasticNet (EN), Gaussian process regression (GPR), Support Vector Machine (SVM), Random Forest (RF), and eXtreme gradient boosting (XGBOOST) [17]. Some of the seminal works include: (a) tree-based model development for predicting methane yield for anaerobic co-digestion for a diverse organic waste stream based on long-term data [18], (b) prediction of biogas yield based on genetic abundance data [19], and (c) data-driven inverse interpretable ML modelling to predict biogas yields [20]. An extensive overview of ML modelling for AD can be found elsewhere [13, 16].

Despite extensive efforts to develop interpretable ML models for predicting methane yields for AD processes without feedstock pretreatment, relevant ML modelling accounting for feedstock pretreatment is relatively sparse. Previous efforts include ML modelling for (a) AD of activated sludge with hydrothermal pretreatment [21], (b) generalizable AD modelling for a range of pretreatment methods (*e.g.*, chemical, ultrasonic, and thermal) of sewage sludge [22], and (c) mechanical grinding and Fe_3O_4 additive-assisted AD of *Arachis hypogea* (*i.e.*, peanut) shells [23]. To our knowledge, there has not been any effort toward developing an optimal ML model selection pipeline for MW-AD process.

The development of ML models for MW-AD of FW as the feedstock adds significant value to the literature from a process modelling and optimization perspective. Specifically, accurate MW-AD process modelling has the potential to facilitate the implementation and practical design of the process towards greater efficiency and sustainability. FW being one of the ubiquitous feedstocks for AD and MW-based pretreatment of feedstock offering efficient and rapid heating has the potential to decarbonize the overall carbon footprint of the biogas production process. This work develops and compares a series of ML models (linear, non-linear, and ensemble-based) to predict methane production based on FW composition, AD conditions, and MW pretreatment parameters. The models are built upon and validated, which upon optimization achieve high accuracy and enhanced interpretability (*i.e.*, via permutation feature importance).

2 Methodology

2.1 Data assimilation

In total, 53 datasets were collected from the literature to develop the data-driven models [24-32]. This included a wide variety of food waste streams (e.g., kitchen waste, organic fraction of municipal solid waste), mono- or co-digestion, thermophilic or mesophilic conditions, and mostly batched reactors. The collected datasets contained a range of information on feedstock properties such as substrate compositions (protein (%P), carbohydrate (%C), lipids (%L)), volatile solids (VS, wt.%), AD reactor operating temperature (°C), hydraulic retention time (HRT, days), pH, reactor volume (L), MW pretreatment temperature (°C), MW pretreatment time (minutes), and methane yield from AD (mL/g VS). The first ten variables (%P, %C, %L, VS, AD temperature, HRT, pH, volume, MW temperature, and MW time) are considered the predictor variables. In contrast, the methane yield is taken as the predicted variable. The raw dataset is provided in the Supplementary Material.

2.2 Data preprocessing methodologies

Since the dataset contains experimental datasets from several different research groups; the assimilated dataset will contain missing values, outliers, and values with dissimilar ranges. This will cause consistency issues while training ML-based continuous regression models, thus affecting their accuracy in predicting methane yield. This problem was addressed by imputing the missing values of an attribute to its corresponding mean [33], ultimately resulting in a complete dataset. It is important to note that these artificially imputed mean values were only performed during the model training and therefore would not affect the model testing.

The constructed dataset would also contain outliers, which require additional preprocessing steps to remove them. Two such popular outlier removal methods such as (a) Z-score normalization and (b) interquartile range (IQR) are considered [4]. The first maps the dataset in terms of the standard normal variate $Z = (X - \mu)/\sigma$, where X is the attribute of interest, μ and σ are the mean and standard deviation of the attribute, respectively. In this case, any datasets with Z scores beyond ±3 are eliminated from the datasets. As a competing method, IQR-based outlier removal removes any datapoint beyond the 25th and 75th percentile.

Following the outlier removal, the dataset was normalized to ensure that the features were appropriately scaled for the ML model development. Two types of normalization were explored (a) max-min normalization (MMN) and (b) maximum absolute scaling (MAS) [33]. MMN uses the transformation function $X' = (X - X_{min})/(X_{max} - X_{min})$ where X_{max} and

X_{min} are the maximum and minimum values of the attribute X, respectively. In contrast, the MAS scales the entire dataset using the absolute maxima of the attribute, *i.e.*, $X' = X/|X_{max}|$.

2.3 Machine learning models

Based on the pre-processed datasets a total of eight different types of ML models are developed and compared, which uses 10 input attributes to predict the methane yield of MW-pretreated AD process. The entire ML workflow has been constructed in Python using the *scikit-learn* library. The pre-processed dataset is split into 80% training and 20% testing fractions to evaluate the model performances. Each of the model was trained using *k*-fold cross validation approach, which ensures high generalizability of the model and mitigate overfitting. The *k*-fold cross-validation was coupled with a hyperparameter optimization engine (i.e., *GridSearchCV* in *scikit-learn*), where initially k = 5 was assigned. The optimization routine heuristically searches through a dictionary of hyperparameters for each model adhering to the *k*-fold cross-validation routine and maximizes the model prediction accuracy. The data-driven modelling pipeline integrated with dataset preprocessing methods are shown in Figure 1. The ML models are described below.

Among the linear ML models, LR and EN are considered. LR can embed several independent variables into the model to predict an output variable (*i.e.*, methane yield). Training an LR model involves determination of unknown regression constants by minimizing the prediction error. The EN is a more sophisticated version of the LR which uses regularisation to mitigate drawbacks of LR. This is achieved via combining the penalty terms of Lasso (L1) and Ridge (L2) regression methods, enabling the model to simultaneously perform variable selection and handle correlated predictors. This becomes important when the datasets involve a larger number (*i.e.*, 10+) of input attributes.

From the pool of non-linear models, ANN, KNN, SVM, and GPR have been selected. Multilayer perceptron (MLP)-based ANN is considered due to its deep non-linear pattern recognition abilities from complex physical datasets. The key to develop an MLP-based ANN is identifying the optimal number of neurons, hidden layer, weights, biases, and activation function. To determine an optimal combination of these hyperparameters for a certain dataset, ANNs must therefore be trained using an hyperparameter optimization engine. KNN model predicts output variables based on individual datapoints and its proximity to *k* neighbouring datapoint. The number of *k* instances in the training dataset is usually determined using statistical distance from the data cluster centroid with Euclidean or Manhattan distances. These further embed onto weighted averaging that determines the influence of neighbouring points

on predicting a target variable. The SVM model is a non-parametric, non-probabilistic method which are suitable for high dimensional datasets handling large number of input/output variables. The model maps input features into a multi-dimensional space using non-linear kernel function, further creating an optimal hyperplane to differentiate between various subsets. In contrast, GPR is a Bayesian probabilistic regression method beneficial for datasets with high variances. The GPR method determines covariance of model predictions which enables uncertainty quantification, generally overlooked by the other ML models.

Among ensembled tree models, RF and XGBoost are chosen due to their complex data learning capabilities for regression applications. Both these models combine many decision trees via ensembling, which ultimately mitigate overfitting issues. The RF is a bagging technique where each tree is trained on a random subset of the training dataset. These individual predictions are then unified via statistical metrics (*e.g.*, mean, median, and mode) towards a robust final prediction, ultimately increasing the model generalizability. XGBoost, on the other hand, is a boosting-based ensembled learning methods where deeper trees are grown in an additive manner. It implements a boosting framework that bases predictions on individual decision trees while simultaneously mitigating errors introduced from each tree. Features such as regularization and randomization minimize the loss function, resulting in reduced overfitting. In general, it is important to note that boosting-based algorithms have shorter training time that bagging algorithms.

2.4 Model performance and interpretability

The root mean squared error (RMSE) and coefficient of determination (R^2) are considered performance metrics for the ML-base regression models.

$$R^2 = \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}} \quad (2)$$

Here y_i and $\hat{y}$ are the true and predicted values of the output attribute (*i.e.*, methane yield), respectively; $\bar{y}$ is the mean of the methane yields, and N is the total number of datasets, which is 53.

In addition, understanding the dependence of model predictions on the input features (i.e., model interpretability) is essential. Being a global interpretability analysis method, permutation feature importance is chosen that provide an overall correlation strength for each predictor variable toward methane yield prediction. This technique is particularly useful for non-linear

or opaque estimators and involves randomly shuffling the values of a single feature and observing the resulting degradation of the model's accuracy. By disrupting the relationship between the predictor and the predicted, it is determined how much a model relies on that predictor. It is important to note that PFI is a model-agnostic (*i.e.*, model-independent) method.

3 Results and Discussion

3.1 Statistical analysis of the dataset

To understand the correlations between variables in the assimilated dataset, which substantiate the physics of MW-AD process, a preliminary statistical analysis is carried out. This includes exploratory analysis on all the variables, data spread visualization, and correlation quantification (see Figure 2). Coupling MW pretreatment with AD increases the digestibility of organics by effective decomposition of extracellular polymeric substances (*e.g.*, protein, carbohydrate), which would then be easily available to microbial communities. The substrate concentration, reactor operating conditions, and MW conditions altogether regulate the methane yield as suggested by the exploratory data analysis (see Figure 2A). To understand the linear correlation strength of any two variables in the dataset, the Pearson Correlation Coefficient (*PCC*) is evaluated. *PCC* ~ ±1 signifies that the variables are highly correlated, while a $PCC = 0$ suggests that the attributes are uncorrelated. The *PCC* between any two attributes x_i and y_i is defined as,

$$PCC = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3)$$

The *PCC*s are shown in Figure 2C via a two-dimensional map where the diameters of the circles are proportional to the *PCC* values. This reveals that the methane yield is positively correlated with the pH, lipid content, and microwave conditions (*i.e.*, time and temperature). In contrast, negative correlation was observed between the protein and carbohydrate contents, VS, AD temperature and HRT.

3.2 Systematic optimization of the ML models

Following the statistical analysis, a range of *what-if* scenarios were investigated for developing an optimal ML model selection pipeline from a pool of eight different models (LR, EN, GPR, KNN, SVM, ANN, RF, and XGBoost). Figure 3 shows the effects of applying different data preprocessing (*i.e.*, outlier removal and normalization), dimension reduction (*i.e.*, principal

component analysis (PCA)), and hyperparameter optimization methods. As mentioned above, the R^2 and RMSE values are used for the accuracy quantification of the ML models.

A high-level comparison across Figures 3A and 3B reveals that the ML models developed using Z-score-based outlier removal methods provide $R^2 \sim 0.92$ with RMSE ~ 38.5 mL/gVS, where RF, KNN, and XGBoost outperform the other models. In contrast, the IQR-coupled ML models fail to predict the methane yield accurately, thus providing unrealistic R^2 values. This is attributed to the fact that IQR is extremely sensitive to dataset removal that removes any data points outside the 25^{th} and 75^{th} quartile. Inspecting Figure 2B suggests that for the present dataset, many datapoints are beyond this range, which makes the IQR method unfavorable. In contrast, the *Z*-score-based outlier detection is much more conservative in removing outliers, relying on μ and $\pm 3\sigma$ values. After selecting the optimal outlier removal method, the effect of utilizing two different data normalization methods (MMN and MAS) on the model performance is explored. Figures 3C and 3D suggest that either of the normalizations can provide accurate model development. The highest accuracy was observed with the ANN model achieving R^2 values up to 0.94, with RMSE as low as 33.5 mL/gVS. Based on this analysis, the Z-score outlier removal with MMN was used for all subsequent analyses.

Coupling dimensionality reduction methods (e.g., PCA) with ML models helps toward feature engineering, eliminates collinearity, and can prevent model overfitting. To understand if PCA is required for the current model pipeline development, all the models were integrated with the PCA-based feature reduction method. Inspecting Figure 3E reveals that although R^2 and RMSE values for some ML models improve when coupled with PCA, it does not drastically change their values. The KNN model outperforms other methods, with an $R^2 \sim 0.92$ and RMSE ~ 38 mL/gVS. The potential reason for not gaining additional accuracy improvement by adding PCA might be attributed to the size of the dataset, where the current dataset is at least an order of magnitude smaller than the scenarios where PCA can provide better results. Subsequently, the ML models were subjected to a 5-fold cross-validation routine with an automatic hyperparameter optimization algorithm (*i.e.*, *GridSearchCV*). The cross-validation coupled with hyperparameter mitigates model overfitting, provides a generic model accuracy averaged over multiple trials, and ensures model generalizability for unseen (*i.e.*, testing) datasets. The optimal setting of hyperparameters for each ML model is provided in Table 1. Figure 3F shows that the SVM model has the highest predictive accuracy after hyperparameter optimization, with an $R^2 \sim 0.84$ and RMSE ~ 33.5 mL/gVS.

3.3 Performance of optimal ML models

The accuracy of methane yield prediction across eight different ML models is visualized in the parity plots shown in Figure 4. The dotted lines represent the ideal prediction line, with an optimal model aligning predicted values closely to this line. Among the linear models (Figures 4a and 4b), the LR and EN models achieved training R2 values of 0.78 and 0.77, respectively, with corresponding RMSE values of 57.64 and 59.24 mL/gVS. For the testing phase, the LR model retained an R2 of 0.72 and RMSE of 71.13 mL/gVS, whereas the EN model exhibited a slight performance drop with an R2 of 0.67 and RMSE of 76.92 mL/gVS. The smaller difference between training and testing accuracies in the LR model suggests better generalization ability. This may be because EN incorporates regularization parameters, which, while beneficial for preventing overfitting, require larger datasets for optimal tuning and effective performance.

Despite the acceptable performance of linear models, AD is governed by complex biokinetic interactions that involve non-linear relationships between operational and compositional parameters [13]. Hence, non-linear ML models are expected to provide superior predictive capabilities for methane yield.

Figures 4c–4f present the predictive performance of non-linear models, including SVM, KNN, ANN, and GPR. These models demonstrated significantly improved accuracy, with training R2 values of 0.94, 1.0, 0.97, and 0.96, and RMSE values of 29.8, 5.14, 21.4, and 24.5 mL/gVS, respectively. In the testing phase, these models retained R2 values of 0.94, 0.92, 0.93, and 0.92, with RMSE values of 33.98, 37.23, 36.06, and 37.59 mL/gVS, respectively. These RMSE values, being within 10% of the mean methane yield, indicate that the developed ML pipeline can effectively predict AD performance trends. Similar observations have been reported in prior studies, where ANN-based models outperformed linear regressors when predicting biogas yields from pretreated lignocellulosic and food waste substrates [17].

Ensemble models such as RF and XGBoost exhibited the highest accuracy during training, with R2 values of 0.96 and 0.99 and RMSE values of 25.52 and 14.04 mL/gVS, respectively (Figures 4g and 4h). However, their testing performance revealed increased RMSE values of 36.98 mL/gVS (RF) and 41.16 mL/gVS (XGBoost), suggesting overfitting. This aligns with findings with literature [16], where ensemble-based models, while powerful, often struggle with generalization when trained on small datasets due to their high sensitivity to outliers and redundant variables.

Although non-linear and ensemble models demonstrated superior predictive power, they also showed a tendency to overfit, particularly for KNN, ANN, GPR, XGBoost, and RF models. The SVM model, however, balanced training and testing accuracy effectively, with relatively low RMSE values, making it a robust choice for methane yield prediction. The overfitting observed in other models is likely due to the limited dataset size (53 entries), which restricts their ability to generalize across different feedstock conditions. Previous studies have reported that larger datasets (>200 entries) significantly improve the performance of ANN and ensemble-based models by allowing them to better capture the non-linear biokinetics of AD [16, 21].

These findings highlight the need for a carefully curated dataset to enhance ML model robustness for methane yield prediction in MW-assisted AD systems. While MW pretreatment plays a crucial role in solubilizing organic matter, the variability in feedstock composition and process parameters necessitates advanced ML approaches that effectively balance accuracy and generalizability.

3.4 Model-agnostic global feature importance analysis

To elucidate the relative importance of various predictor variables in forecasting methane yield during, a feature importance analysis was conducted using PFI, a global interpretability method (Figure 5). Analysis indicated that pH was the most influential factor affecting methane yield in MW-assisted AD, followed by lipid and carbohydrate compositions. The methanogenesis stage of AD is highly sensitive to pH fluctuations, with an optimal range of approximately 6.8–7.2. Deviations from this range can adversely affect microbial activity and process stability. MW pretreatment alters the chemical composition of substrates by solubilizing complex biopolymers, enhancing biodegradability, and releasing by-products like organic acids, leading to decreased pH. Studies have shown that MW pretreatment can increase organic matter solubilization, thereby improving methane production [9]. Interestingly, fluctuations in feedstock pH during AD have a more pronounced impact on methane yield than the operational parameters associated with MW pretreatment. This suggests that unless MW pretreatment is applied under extreme conditions, its influence on methane yield is secondary to factors such as pH and substrate composition [34].

Hydrothermal pretreatment, another thermal method for enhancing anaerobic digestibility, involves exposing substrates to high temperatures (120–220°C) under pressurized conditions, leading to extensive breakdown of complex organic matter. However, this method can produce

inhibitory compounds like furfurals and hydroxymethylfurfural (HMF), which may suppress microbial activity if not properly managed [35]. In contrast, MW pretreatment utilizes rapid, selective heating through dielectric polarization, targeting polar molecules such as water. This leads to localized overheating, promoting cell wall disruption and release of intracellular components without significantly degrading sugars into inhibitory compounds [27, 29]. Consequently, MW pretreatment enhances bioavailability while minimizing the risk of toxic by-product formation. This distinction aligns with previous studies suggesting that variations in feedstock composition have a greater influence on methane yield than changes in pretreatment conditions. For instance, studies that maintained constant MW pretreatment parameters while altering feedstock chemical composition observed more significant deviations in methane yield compared to those that modified MW pretreatment conditions alone [36].

While controlling MW pretreatment conditions can influence methane yield, the effect is relatively moderate unless extreme MW treatment settings are applied. This emphasizes the need for tailored modelling strategies that prioritize microbial and biochemical parameters over purely physical pretreatment variables. Future research should explore integrating advanced multi-omics data with machine learning approaches to better capture the microbial dynamics governing AD performance under different pretreatment strategies.

4 Conclusions

To facilitate data-driven process optimization of MW-pretreated AD of FW, the work herein developed and compared a series of ML models i.e., linear, non-linear, and ensembled-learning models. The predictor variables included information on FW composition, AD reactor conditions, and MW pretreatment parameters. Upon systematic comparison of the selection of data preprocessing techniques, cross-validation, and hyperparameter optimization, models achieved excellent accuracy in predicting the methane yield for MW-pretreated AD of FW. The optimized SVM-based model coupled with the Z-score method as outlier removal and the Max-Min normalization technique provided R^2 values in the range of 0.85-0.9 with an RMSE of 34 mL/gVS (representing less than 10% relative error). The model's interpretability was augmented by permutation feature importance analysis, a global model-agnostic model explainer. It projected insights into the most influential variables that regulate methane yield for MW-AD processes, suggesting that AD reactor pH and FW compositions were more influential than MW operational parameters. The developed model with added experimental datasets, in the future, could be used for what-if scenario analysis, life cycle assessment

framework, and reactor control frameworks towards rapid process optimization. This will ultimately facilitate the practical application of AD-based waste valorization systems and contribute to a circular economy.

Acknowledgments

Siming You and William T. Sloan acknowledges the financial support from the Engineering and Physical Sciences Research Council (EPSRC) Programme Grant (EP/V030515/1). Siming You would also like to acknowledge the financial support from the Royal Society International Exchange Scheme (EC\NSFC\211175). Rohit Gupta acknowledges the Royal Society Newton International Fellowship (NIF\R1\211013). All data supporting this study are provided in full in the paper and supplementary material.

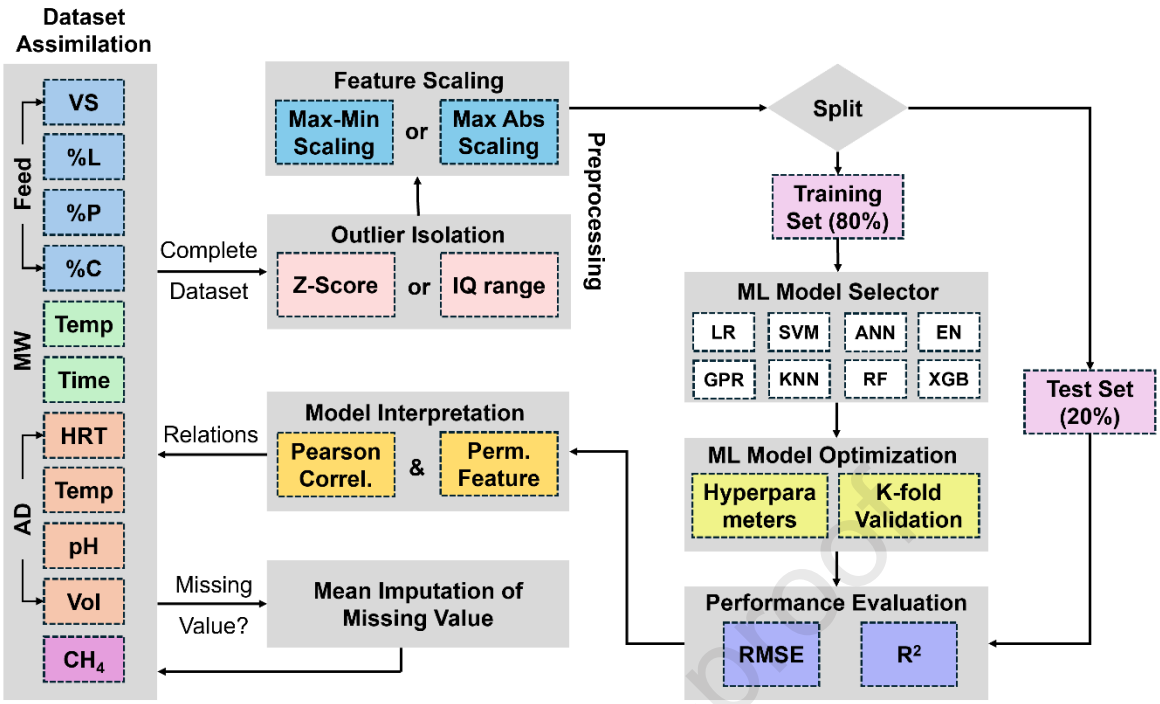


Figure 1. Sequential stages of the machine learning model development to predict methane yield. Following on to the preliminary dataset construction, missing values in the dataset were imputed with respective means. The dataset was then subjected data preprocessing that included outlier removal and variables scaling. The pre-processed dataset was split into training and testing sets using which a range of ML models were constructed. The predictive accuracy of the optimized ML model was quantified in terms of RMSE and R^2 metrics. Finally, the relationships between the variables were understood via Permutation Feature Importance analysis and Pearson Correlation Coefficient.

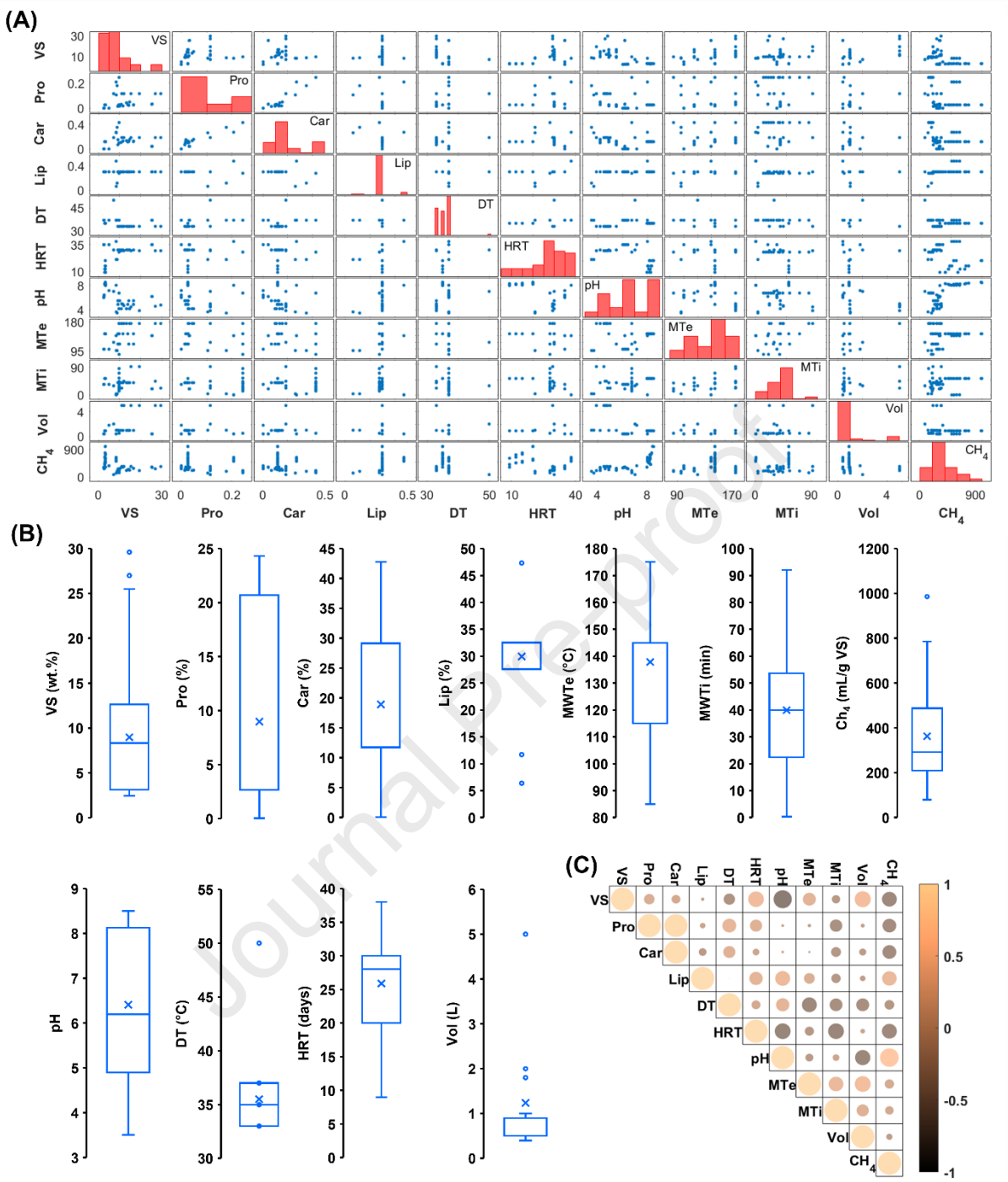


Figure 2. Statistical analysis of the assimilated dataset. (A) Exploratory data analysis across different variables via two-ways plots. (B) Box-whisker plot showing spread of different variables. (C) Pearson correlation coefficient map across any two variables where the diameter of the circles is proportional to the correlation coefficient. VS: Volatile Solids, Pro: Protein, Car: Carbohydrate, Lip: Lipid, DT: Digester temperature, HRT: Hydraulic retention time, MWTe: Microwave pretreatment temperature, MWTi: Microwave pretreatment time, Vol: Digester volume, CH_4: Methane yield.

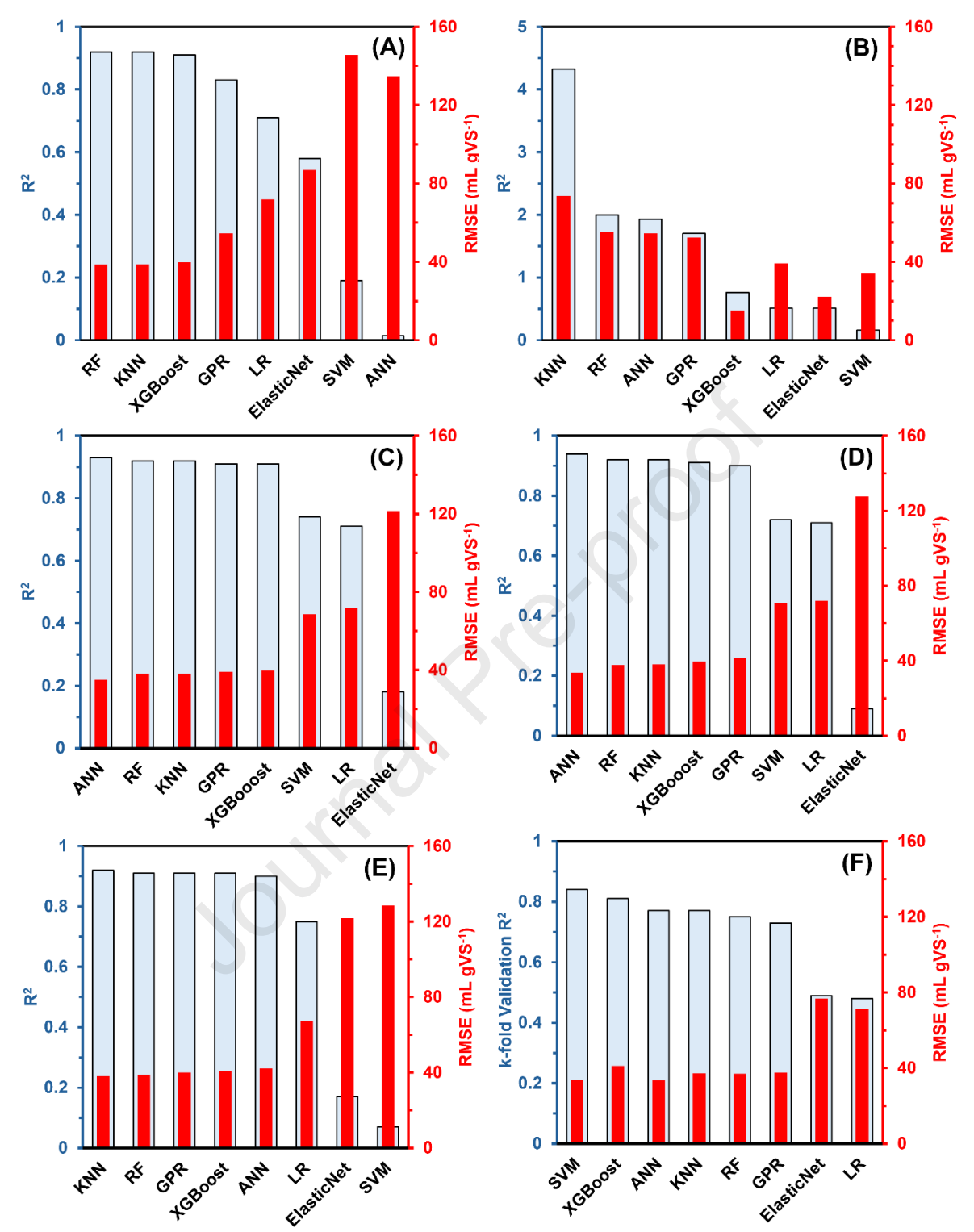


Figure 3. Performance assessment of different data-driven models using R^2 (light blue) and RMSE (red). (A) Z-score based outlier removal, (B) interquartile range-based outlier removal, (C) max-min normalization, (D) max absolute scaling, (E) with principal component analysis, (F) after hyperparameter optimization.

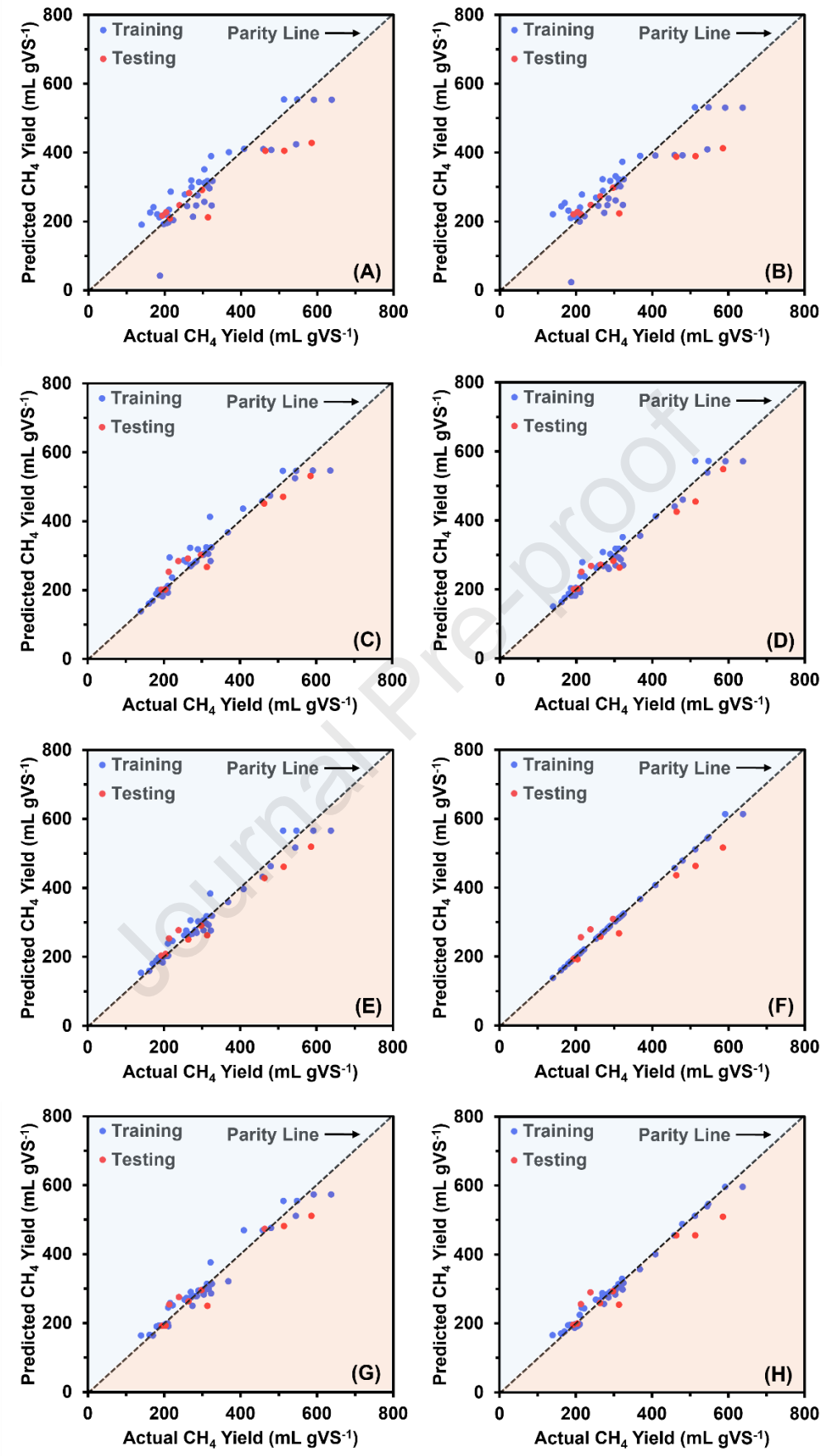


Figure 4. Parity plots obtained after optimizing different ML models. (A) LR, (B) ElasticNet, (C) SVM, (D) ANN, (E) GPR, (F) KNN, (G) RF, and (H) XGBoost.

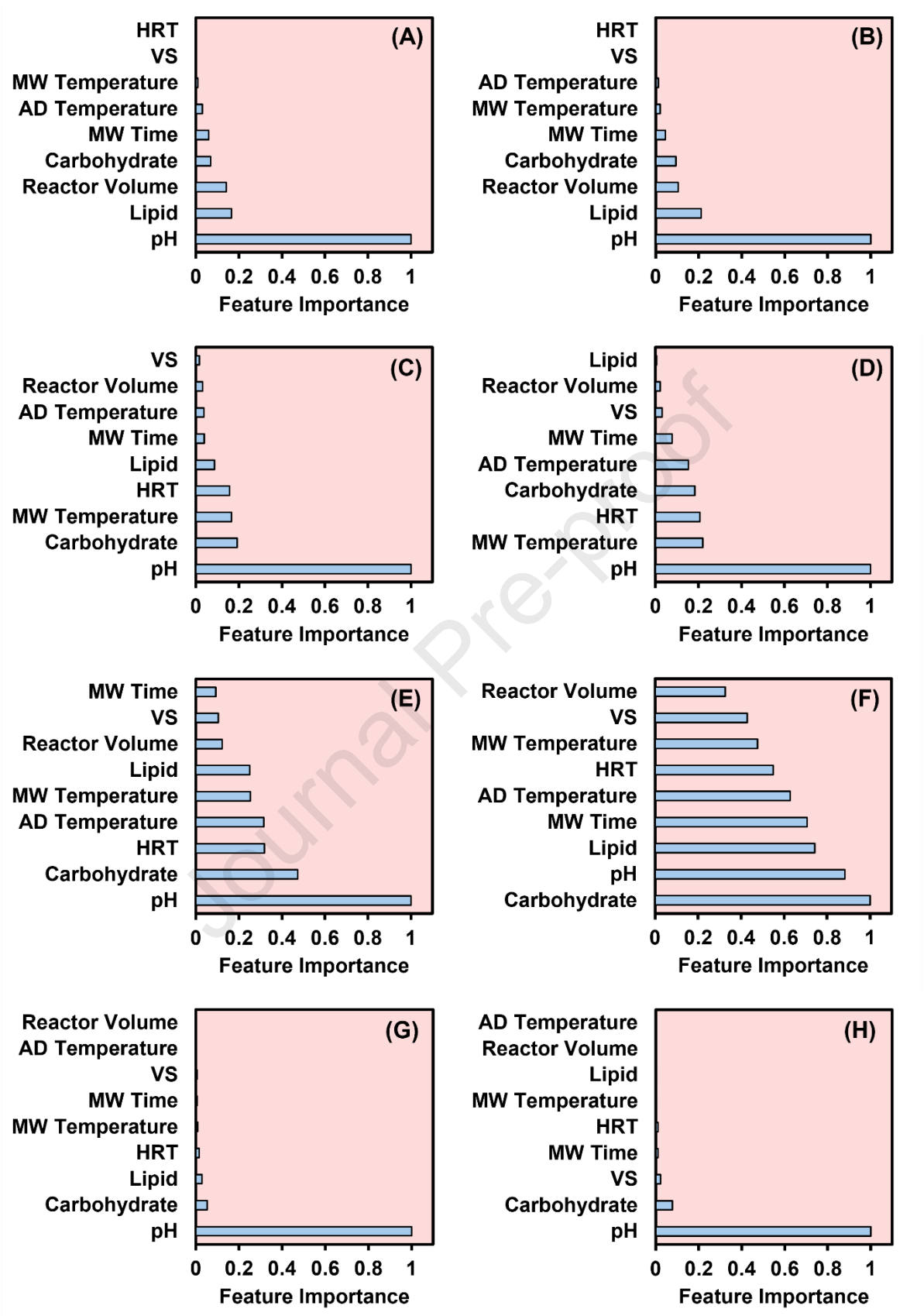


Figure 5. Premutation feature importance (normalized) analysis showing relative importance of predictor variables for different ML models after optimization. (A) LR, (B) ElasticNet, (C) SVM, (D) ANN, (E) GPR, (F) KNN, (G) RF, and (H) XGBoost. The absence of protein content in these plots is due to its exclusion during ML model development

Table 1. Optimal hyperparameter values of the ML models using *GridSearchCV* algorithm.

ML Model	Optimal Hyperparameter Combination
Linear Regression	Fit Intercept: False
ElasticNet	Fit Intercept: False, α: 0.1, L1 Ratio: 0.9
Support Vector Machine	C: 50, ε: 0.1, Kernel Type: Polynomial
K-Nearest Neighbour	No. Neighbours: 9, Weight Function: Distance
Artificial Neural Network	Hidden Layer Size: 100, Activation Function: Logistic, Solver: SGD, Max Iterations: 1000
Gaussian Process Regression	Kernel Type: RBF 1, Normalise: True
Random Forest	No. of Trees: 50, Max Depth of Trees: 5, Min Leaf Samples: 2, Min Split Samples: 2
eXtreme Gradient Boosting	No. of Boosting rounds: 50, Max Depth of Trees: 3, Learning Rate: 0.1, Subsample Ratio 1: 0.8, Subsample Ratio 2: 1

References

1. B. Alves, *Global waste generation-statistics & facts.* Statista, 2023.
2. P. Roy, A.K. Mohanty, P. Dick, and M. Misra, *A review on the challenges and choices for food waste valorization: Environmental and economic impacts.* ACS environmental Au, 2023. **3**(2): p. 58-75.
3. R. Gupta, R. Miller, W. Sloan, and S. You, *Economic and environmental assessment of organic waste to biomethane conversion.* Bioresource Technology, 2022. **345**: p. 126500.
4. Z.H. Ouderji, R. Gupta, A. Mckeown, Z. Yu, C. Smith, W. Sloan, and S. You, *Integration of anaerobic digestion with heat Pump: Machine learning-based technical and environmental assessment.* Bioresource Technology, 2023. **369**: p. 128485.
5. W. Li, R. Gupta, Z. Zhang, L. Cao, Y. Li, P.L. Show, V.K. Gupta, S. Kumar, K.-Y.A. Lin, and S. Varjani, *A review of high-solid anaerobic digestion (HSAD): From transport phenomena to process design.* Renewable and Sustainable Energy Reviews, 2023. **180**: p. 113305.
6. H. Carrere, G. Antonopoulou, R. Affes, F. Passos, A. Battimelli, G. Lyberatos, and I. Ferrer, *Review of feedstock pretreatment strategies for improved anaerobic digestion: From lab-scale research to full-scale application.* Bioresource technology, 2016. **199**: p. 386-397.
7. M. Atelge, A. Atabani, J.R. Banu, D. Krisa, M. Kaya, C. Eskicioglu, G. Kumar, C. Lee, Y. Yildiz, and S. Unalan, *A critical review of pretreatment technologies to enhance anaerobic digestion and energy recovery.* Fuel, 2020. **270**: p. 117494.
8. B. Ahmed, V.K. Tyagi, K. Aboudi, A. Naseem, C.J. Álvarez-Gallego, L.A. Fernández-Güelfo, A.A. Kazmi, and L.I. Romero-García, *Thermally enhanced solubilization and anaerobic digestion of organic fraction of municipal solid waste.* Chemosphere, 2021. **282**: p. 131136.
9. S. Simonetti, C. Fernández Martín, and D. Dionisi, *Microwave pre-treatment of model food waste to produce short chain organic acids and ethanol via anaerobic fermentation.* Processes, 2022. **10**(6): p. 1176.
10. Y. Li, L.C. Campos, and Y. Hu, *Microwave pretreatment of wastewater sludge technology—a scientometric-based review.* Environmental Science and Pollution Research, 2024. **31**(18): p. 26432-26451.
11. F.-M. Pellera and E. Gidarakos, *Microwave pretreatment of lignocellulosic agroindustrial waste for methane production.* Journal of environmental chemical engineering, 2017. **5**(1): p. 352-365.
12. J. Ariunbaatar, A. Panico, G. Esposito, F. Pirozzi, and P.N. Lens, *Pretreatment methods to enhance anaerobic digestion of organic solid waste.* Applied energy, 2014. **123**: p. 143-156.
13. R. Gupta, L. Zhang, J. Hou, Z. Zhang, H. Liu, S. You, Y.S. Ok, and W. Li, *Review of explainable machine learning for anaerobic digestion.* Bioresource technology, 2023. **369**: p. 128468.

14. D.J. Batstone, J. Keller, I. Angelidaki, S. Kalyuzhnyi, S. Pavlostathis, A. Rozzi, W. Sanders, H. Siegrist, and V. Vavilin, *The IWA anaerobic digestion model no 1 (ADM1).* Water Science and technology, 2002. **45**(10): p. 65-73.

15. R. Mo, W. Guo, D. Batstone, J. Makinia, and Y. Li, *Modifications to the anaerobic digestion model no. 1 (ADM1) for enhanced understanding and application of the anaerobic treatment processes–A comprehensive review.* Water Research, 2023: p. 120504.

16. I.A. Cruz, W. Chuenchart, F. Long, K. Surendra, L.R.S. Andrade, M. Bilal, H. Liu, R.T. Figueiredo, S.K. Khanal, and L.F.R. Ferreira, *Application of machine learning in anaerobic digestion: Perspectives and challenges.* Bioresource Technology, 2022. **345**: p. 126433.

17. R. Gupta, Z.H. Ouderji, Uzma, Z. Yu, W.T. Sloan, and S. You, *Machine learning for sustainable organic waste treatment: a critical review.* npj Materials Sustainability, 2024. **2**(1): p. 5.

18. Y. Wang, T. Huntington, and C.D. Scown, *Tree-based automated machine learning to predict biogas production for anaerobic co-digestion of organic waste.* ACS Sustainable Chemistry & Engineering, 2021. **9**(38): p. 12990-13000.

19. F. Long, L. Wang, W. Cai, K. Lesnik, and H. Liu, *Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data.* Water Research, 2021. **199**: p. 117182.

20. J. Li, L. Zhang, C. Li, H. Tian, J. Ning, J. Zhang, Y.W. Tong, and X. Wang, *Data-driven based in-depth interpretation and inverse design of anaerobic digestion for CH4-rich biogas production.* ACS ES&T Engineering, 2022. **2**(4): p. 642-652.

21. M.G. Fard and E.H. Koupaie, *Machine learning assisted modelling of anaerobic digestion of waste activated sludge coupled with hydrothermal pre-treatment.* Bioresource Technology, 2024. **394**: p. 130255.

22. X. Cheng, R. Xu, Y. Wu, B. Tang, Y. Luo, W. Huang, F. Wang, S. Fang, Q. Feng, and Y. Cheng, *Predicting and evaluating different pretreatment methods on methane production from sludge anaerobic digestion via automated machine learning with ensembled semisupervised learning.* ACS ES&T Engineering, 2023. **4**(3): p. 525-539.

23. K.O. Olatunji, D.M. Madyira, N.A. Ahmed, O. Adeleke, and O. Ogunkunle, *Modeling the biogas and methane yield from anaerobic digestion of Arachis hypogea shells with combined pretreatment techniques using machine learning approaches.* Waste and Biomass Valorization, 2023. **14**(4): p. 1123-1141.

24. J. Liu, M. Zhao, C. Lv, and P. Yue, *The effect of microwave pretreatment on anaerobic co-digestion of sludge and food waste: Performance, kinetics and energy recovery.* Environmental Research, 2020. **189**: p. 109856.

25. J. Marin, K.J. Kennedy, and C. Eskicioglu, *Effect of microwave irradiation on anaerobic degradability of model kitchen waste.* Waste Management, 2010. **30**(10): p. 1772-1779.

26. B. Deepanraj, V. Sivasubramanian, and S. Jayaraj, *Effect of substrate pretreatment on biogas production through anaerobic digestion of food waste.* International Journal of Hydrogen Energy, 2017. **42**(42): p. 26522-26528.

27. H. Shahriari, M. Warith, M. Hamoda, and K. Kennedy, *Evaluation of single vs. staged mesophilic anaerobic digestion of kitchen waste with and without microwave pretreatment.* Journal of Environmental Management, 2013. **125**: p. 74-84.

28. I. Pecorini, F. Baldi, E.A. Carnevale, and A. Corti, *Biochemical methane potential tests of different autoclaved and microwaved lignocellulosic organic fractions of municipal solid waste.* Waste Management, 2016. **56**: p. 143-150.

29. J. Zhang, C. Lv, J. Tong, J. Liu, J. Liu, D. Yu, Y. Wang, M. Chen, and Y. Wei, *Optimization and microbial community analysis of anaerobic co-digestion of food waste and sewage sludge based on microwave pretreatment.* Bioresource Technology, 2016. **200**: p. 253-261.

30. H. Shahriari, M. Warith, M. Hamoda, and K.J. Kennedy, *Anaerobic digestion of organic fraction of municipal solid waste combining two pretreatment modalities, high temperature microwave and hydrogen peroxide.* Waste Management, 2012. **32**(1): p. 41-52.

31. M.V.T. Suruagy, A.B. Ross, and A. Babatunde, *Influence of microwave temperature and power on the biomethanation of food waste under mesophilic anaerobic conditions.* Journal of Environmental Management, 2023. **341**: p. 117900.

32. L. Yue, J. Cheng, S. Tang, X. An, J. Hua, H. Dong, and J. Zhou, *Ultrasound and microwave pretreatments promote methane production potential and energy conversion during anaerobic digestion of lipid and food wastes.* Energy, 2021. **228**: p. 120525.

33. S. Ascher, W. Sloan, I. Watson, and S. You, *A comprehensive artificial neural network model for gasification process prediction.* Applied Energy, 2022. **320**: p. 119289.

34. X. Kan, J. Zhang, Y.W. Tong, and C.-H. Wang, *Overall evaluation of microwave-assisted alkali pretreatment for enhancement of biomethane production from brewers' spent grain.* Energy Conversion and Management, 2018. **158**: p. 315-326.

35. F. Passos, J. Carretero, and I. Ferrer, *Comparing pretreatment methods for improving microalgae anaerobic digestion: thermal, hydrothermal, microwave and ultrasound.* Chemical Engineering Journal, 2015. **279**: p. 667-672.

36. L. Wang, F. Long, W. Liao, and H. Liu, *Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms.* Bioresource technology, 2020. **298**: p. 122495.

Highlights

- Data-driven models predict methane production for Microwave-pretreated AD
- A suite of data-preprocessing and ML modelling techniques were explored
- The ML models were either linear, non-linear, and ensemble-based
- Support vector machine has highest accuracy (R^2 = 0.9 with <10% relative error)
- Feature analysis identified pH and food waste composition as high-importance parameters

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: