ENVIRONMENTAL RESEARCH
COMMUNICATIONS

INCLUSIVE PUBLISHING TRUSTED SCIENCE

Part of the ENVIRONMENTAL RESEARCH SERIES

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Predicting biogas production in real scale anaerobic digester under dynamic conditions with machine learning approach

To cite this article: M. Erdem Isenkul *et al* 2025 *Environ. Res. Commun.* **7** 065016

Environmental Research Communications

PAPER

Predicting biogas production in real scale anaerobic digester under dynamic conditions with machine learning approach

M. Erdem Isenkul[1,*], Sevgi Güneş-Durak[2], Yasemin Poyraz Kocak[3], İnci Pir[4], Mertol Tüfekci[5,6,*], Güler Türkoğlu Demirkol[7], Selçuk Sevgen[8], Aslı Seyhan Çığgın[9] and Neşe Tüfekci[7]

[1] Department of Computer Engineering, Faculty of Engineering, Istanbul University-Cerrahpaşa, Avcılar, 34320 Istanbul, Turkey
[2] Department of Environmental Engineering, Faculty of Engineering-Architecture, Nevsehir Haci Bektas Veli University, Nevsehir 50300, Turkey
[3] Department of Computer Programming, Vocational School of Technical Sciences, Istanbul Universitesi—Cerrahpasa, 34850 Istanbul, Turkey
[4] Faculty of Mechanical Engineering, Istanbul Technical University, Gumussuyu, Istanbul 34437, Turkey
[5] Centre for Engineering Research, University of Hertfordshire, College Lane Campus, Hatfield AL10 9AB, United Kingdom
[6] School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, Hertfordshire AL10 9AB, United Kingdom
[7] Department of Environmental Engineering, Istanbul University-Cerrahpaşa, Avcılar, 34320 Istanbul, Turkey
[8] Department of Software Engineering, Faculty of Engineering, Istanbul University-Cerrahpaşa, Avcılar, 34320 Istanbul, Turkey
[9] Department of Environmental Engineering, Faculty of Engineering, Akdeniz University, Dumlupınar Bulvarı, Antalya, 07058, Turkey
[*] Authors to whom any correspondence should be addressed.

E-mail: eisenkul@iuc.edu.tr, yasemin.poyraz@iuc.edu.tr, m.tufekci@herts.ac.uk and sevgens@iuc.edu.tr

Keywords: anaerobic digestion, biogas production, machine learning, support vector regression (SVR), wastewater treatment

OPEN ACCESS

RECEIVED
13 March 2025

REVISED
15 May 2025

ACCEPTED FOR PUBLICATION
3 June 2025

PUBLISHED
13 June 2025

Abstract

Biogas production through anaerobic digestion (AD) of industrial organic waste and wastewater offers a sustainable method for energy recovery. However, since process efficiency heavily relies on operational factors, continuous monitoring of the AD process and the implementation of necessary operational strategies are crucial. In recent years, the use of machine learning techniques (ML) has become widespread for analysing the effects of operational factors on anaerobic digestion efficiency. Among these, Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel has been used to predict biogas yield based on diverse operating parameters. This study aimed to investigate the predictability of changes in biogas production using the SVR algorithm with an RBF kernel in a full-scale anaerobic digester treating wastewater from a fruit processing plant. In the model, biogas production was estimated based on variations in selected operational parameters, achieving a regression coefficient (R^2) of 0.8983 ± 0.03 with mean square error (MSE) of 0.0047 ± 0.0017. The model's performance was evaluated using 10-fold cross-validation techniques and relevant statistical indicators to ensure robustness and generalisability. Hyperparameter tuning was conducted to enhance prediction accuracy while reducing model error. The findings demonstrated that ML-based modelling can serve as a reliable and effective tool to improve biogas production efficiency in wastewater treatment applications. Furthermore, the study highlights the potential of such models to support real-time process control and decision making in anaerobic digestion systems operating under variable industrial conditions.

Introduction

Renewable energy incentives and the transition processes towards a circular economy in the context of climate change have increased the significance of biomass as a raw material (Sherwood, 2020). Anaerobic digestion (AD) is one of the most widely utilised technologies for converting waste and raw biomass into feedstock for energy production. AD technology is effectively used in industries that generate high organic content wastewater, such as fruit processing plants, for simultaneous wastewater treatment and energy recovery. However, the composition of wastewater produced in these industries varies based on the type and quantity of raw material

processed at the facility. This variability in wastewater composition, known as influent parameters, leads to unstable conditions in AD and, consequently, fluctuations in biogas production. Achieving high biogas output with AD relies on managing the sensitive microorganisms involved in converting biomass to methane, which is crucial for process control. In other words, the efficiency of AD particularly depends on regulating operational parameters [1].

Operating parameters that affect the performance of anaerobic microorganisms and the efficiency of biogas production include pH, temperature, organic loading rate (OLR), and carbon/nitrogen ratio, which indicate the nutrient balance necessary for microbial growth. Additionally, alkalinity, hydraulic retention time (HRT), and solids content play important roles [2–4]. OLR determines the maximum biogas production and the extent of chemical oxygen demand (COD) removal. Effluent characteristics, including recirculation practises, significantly influence removal efficiency. Methane production and hydrolysis rates are particularly affected by the recirculation rate. However, substrates with high solid content reduce methane production [5], while lower HRT can enhance biogas yield [6]. Each of these parameters is critically important for evaluating the performance and operational efficiency of a wastewater treatment system.

Optimising operating parameters and process configurations to enhance biogas production through AD is a crucial subject for research. However, experimental studies aimed at optimising various operating parameters across different process configurations can be time-consuming and expensive. Therefore, similar to other biochemical processes, modelling studies are conducted to understand and manage AD. Predicting and optimising biogas production is essential for energy savings and efficiency in wastewater treatment facilities. While kinetic models, such as first-order and the Gompertz equation, estimate parameters like reaction rates based on cumulative methane production data from various biomass types [7], mechanistic models that provide more detailed results are used to comprehend the biochemical processes occurring during AD. The most prominent mechanistic model for predicting cumulative biogas yield and composition in AD, based on biochemical and physicochemical processes, is the Anaerobic Digestion Model No. 1 (ADM1) [8].

Mechanistic models such as ADM1 require long-term data that reflect steady-state conditions due to their comprehensive nature. However, in practical applications, such as controlling real-scale industrial anaerobic digesters that operate under variable conditions and varying feedstock contents, the use of mechanistic models presents challenges due to their computational complexity [9]. In contrast, machine learning approaches do not require detailed biochemical assumptions and can effectively learn patterns from operational data, making them highly suitable for real-time biogas yield optimisation under dynamic industrial conditions.

In recent years, machine learning (ML) models and soft computing techniques have emerged as alternative methods for AD modelling. Various regression analysis methods have been developed to predict biogas production yield and COD removal rates based on operating parameters [10]. Regression analysis is a statistical technique used to model and analyse the relationship between a dependent variable and one or more independent variables [11]. In this approach, the developed model is transformed into a continuous-valued output rather than an output derived from a finite set. In other words, a regression model estimates a continuous-valued multivariate function. Many algorithms have been devised for regression problems based on the relationships among variables, such as Multiple Linear Regression (MLR) [12] for linear variables, Polynomial Regression (PR), Support Vector Regression (SVR), or Decision Tree (DT) [13] for nonlinear variables, as well as Ridge, Lasso, or Elastic Net [14] to mitigate the risk of overfitting, Random Forest, Gradient Boosting, or Artificial Neural Networks (ANN) [15] for large and complex datasets, and finally, Bayesian Regression [16] for uncertainty analysis. However, most of these models have been tested in controlled or limited-scale environments, and their comparative performance in real-scale, high-variability wastewater conditions has not been comprehensively addressed.

In a study investigating the performance of three machine learning techniques- namely ANN, Adaptive Neuro-Fuzzy Inference System (ANFIS), and support vector machine (SVM)- for predicting methane production in landfills where municipal solid waste is stored, it was determined that the SVM model outperformed the other ML models in predicting methane production [17]. Although the SVM model excels in estimating biogas production, it has significant drawbacks, including high computational costs when managing large datasets and an extremely slow training process due to the kernel matrix growing quadratically with data size in large datasets [18]. To address these challenges, SVR, a generalised form of SVM designed for regression problems, is applied to model linear or nonlinear hyperplane variables using error tolerance (ϵ) and kernel functions [19]. The main advantages of SVR are (i) its computational complexity, which is independent of the input domain's dimensionality; (ii) the ability to generalise within input data, enhancing the system's prediction efficiency; (iii) adaptability to current data; and (iv) effectiveness in predicting future unknown data [19, 20]. Consequently, SVR exhibits a strong capability to manage complex and nonlinear relationships among various dependent and independent parameters [21]. Its application is becoming increasingly widespread to address issues involving both linear and nonlinear correlations across various engineering challenges, including wastewater treatment simulations. In a recent study, the relationship between process parameters (pH,

oxidation–reduction potential, and conductivity) and daily volatile fatty acid production in an anaerobic digester treating sewage sludge was accurately predicted using the ML model constructed with the SVR algorithm [20]. Nevertheless, many of these applications focus on laboratory-scale systems and often lack rigorous hyperparameter tuning procedures, which limits their robustness in industrial applications.

In order to achieve high accuracy estimation for complex data in SVR algorithm applications, the appropriate selection of hyperparameters such as epsilon, box restriction (C), kernel scale and kernel function is of great importance [22]. In SVR models, especially in order to effectively solve nonlinear and multidimensional problems with limited samples, the appropriate kernel function selection is required. Linear, radial basis function (RBF), polynomial and sigmoid kernel functions are the most frequently preferred kernel functions [23]. Two SVR algorithms with three kernel functions were successfully implemented to relate process parameters (flow rate, OLR, temperature, and influent solid concentration) to COD removal efficiency in a pilot-scale anaerobic digester treating printing and dyeing wastewater [24]. In the study, linear, RBF, and sigmoid kernel functions were utilized, and it was found that using different kernel functions did not significantly affect the SVR algorithm's performance, although the algorithms using the RBF kernel function demonstrated the lowest error tolerance. The RBF kernel has become the preferred kernel function in SVR modelling for estimating biogas production [22, 25] due to its high efficiency and minimal parameter changes [26].

While numerous studies have applied machine learning techniques to model anaerobic digestion processes, the majority have focused on laboratory or pilot-scale reactors operating under stable and controlled conditions, limiting their practical applicability to industrial systems. In particular, the effects of influent variability—caused by seasonal and process-driven fluctuations in real industrial wastewater streams—have not been sufficiently addressed. Furthermore, previous SVR-based studies often omit a detailed hyperparameter tuning process, which is crucial for enhancing model generalisation and avoiding overfitting in nonlinear, high-dimensional settings. Addressing these limitations, this study develops a robust, data-driven approach for biogas prediction by applying an optimised SVR model to operational data from a full-scale UASB reactor treating variable industrial effluent. This contribution aims to bridge the gap between theoretical model development and real-world implementation in anaerobic digestion systems.

To address the identified research gap, this study implements a tuned SVR model on a real-scale UASB reactor dataset with high influent variability, aiming to improve the generalisability and practical applicability of ML-based biogas prediction. Considering these challenges, this study investigates the predictability of biogas production in a real-scale Upflow Anaerobic Sludge Blanket (UASB) reactor using a SVR model with an RBF kernel. This study specifically explores how variations in operational parameters—such as influent/effluent COD concentrations, soluble COD, volumetric organic loading rate, and hydraulic retention time—affect biogas yield under industrial conditions. Through the integration of cross-validation and hyperparameter tuning, the proposed model provides a reliable framework for guiding real-time operational decisions in large-scale anaerobic digestion processes. The following sections present the materials and methods, experimental findings, and the implications for biogas optimisation in wastewater treatment applications.

Material and methods

Data collection

In modelling studies, operating parameters monitored for three years during the operation of a UASB-type AD reactor for treating wastewater generated in a fruit processing plant with a capacity of 125,000 tons/year were used. Citrus fruits were used in the process. The wastewater characteristics vary depending on the specific type of citrus fruit processed, resulting in dynamic influent compositions throughout the year. These variations are valuable for evaluating the model's robustness under real industrial conditions. The measured values of the parameters selected as input data for developing and evaluating the model, such as influent COD concentration (COD_{in}), effluent COD concentration (COD_{eff}), effluent soluble COD concentration ($sCOD_{eff}$), volumetric organic loading rate (vOLR), and HRT, along with the measured biogas production taken as output, were summarised in table 1. The independent variables used in the model were determined by considering their effects on biogas production. The selected parameters are the most critical operational variables affecting biogas production in the anaerobic digestion process and are supported by the findings of previous studies. In particular, it is widely recognised in the literature that COD concentration, OLR, and HRT are strongly correlated with biogas production [27]. Therefore, no additional feature selection method was applied in the present study; instead, variables representing the key dynamics of the process were directly included in the model.

The large variation in input parameters, particularly COD concentrations, was measured using the procedure defined in the Standard Methods [28]. In this study, a real-scale UASB reactor was utilised, and no

Table 1. Model input and output parameters.

Parameters	Minimum	Maximum	Mean	Standard deviation
		Inputs		
COD_{in} (mg/l)	820.0	18558.0	6859.1	2604.1
COD_{eff} (mg/l)	20.0	6690.0	1787.8	1001.6
$sCOD_{eff}$ (mg/l)	0.0	2340.0	684.6	451.5
vOLR (kg/m^3)	0.90	30.40	10.54	6.54
HRT (day)	3.20	75.60	21.23	16.39
		Output		
Biogas production (m^3/day)	222.0	14317.0	4284.0	2977.1

additional control reactor was installed. This decision aimed to directly model industrial-scale dynamic operating conditions and generate ML-based predictions using real system data. However, comparative analysis with a control reactor in future studies may enhance the validity of the model. The significant fluctuations in input parameters arise from processing different fruit types in various seasons at the facility. Such variations are deemed beneficial for testing the accuracy of the SVR algorithm.

To clarify the structure and limitations of the dataset, it is important to note that the model was developed using a real-scale dataset collected over three years from a single industrial UASB reactor treating citrus fruit processing wastewater. The dataset contains 300 instances, each representing a daily average of key operational parameters and corresponding biogas production. Data were collected using online sensors and lab-based measurements, adhering to standard monitoring protocols. No artificial filtering or interpolation was applied to smooth the data, allowing the model to reflect true industrial variability. Although this dataset provides high temporal resolution and captures seasonal dynamics effectively, it is limited to a single facility and type of wastewater. As such, the generalisability of the findings to other industrial sectors or digester configurations requires validation with diverse datasets. The selected features (CODin, CODeff, sCODeff, OLR, and HRT) were chosen based on both domain knowledge and their consistent availability across the monitoring period.

Support vector regression (SVR)

SVM is a powerful supervised learning algorithm used for binary classification and regression problems, aiming to find the hyperplane that best separates the data by formulating them as convex optimisation problems [29]. The hyperplane is represented in terms of support vectors. SVR is a generalised form of SVM adapted for regression problems and was pioneered by Vapnik and Chervonenkis [19, 30]. The main objective of SVR is to find a function that best fits the training data while ensuring good predictive performance on new, unseen data. SVR also defines an ε-insensitive zone around the function, referred to as the ε-tube. This tube redefines the optimisation problem to identify the tube that most accurately approximates the continuous-valued function while balancing model complexity and prediction error. The continuous-valued function can be expressed as seen in equation (1):

$$y = f(x) = \langle w, x \rangle + b = \sum_{j=1}^{M} w_j x_j + by,\ b\epsilon\Re x,\ w \in \Re^M \tag{1}$$

where M denotes the order of the polynomial used to approximate a function. Additionally, x, y, w, and b denote the input feature vector, target output, weight vector, and bias term, respectively. The magnitude of the weights can be viewed as an indicator of flatness. To achieve a flat function f, the weight vector w must be small. A convex optimisation problem is utilised to minimise the regression risk R_{reg}, as shown in equation (2):

$$R_{reg} = \frac{1}{2}w^2 + C\sum_{i=1}^{N}(\varepsilon_i + \varepsilon_i^*) \tag{2}$$

where the constant $C > 0$ acts as a regularization parameter, referred to as the box constraint. This parameter balances the trade-off between the smoothness of f and the allowable deviation beyond the ε margin. Increasing C imposes a greater penalty on errors, resulting in more precise predictions. However, a higher C also elevates model complexity, which diminishes its generalization ability compared to a model with a lower C value. In this study, the C value is determined to be 0.3556, identified as the optimal value through cross-validation.

SVR interprets function approximation as an optimisation problem that seeks to find the smallest possible tube centred around the surface while minimising the error value, which represents the difference between predicted and actual outputs. SVR employs an ε-insensitive loss function that discards predictions deviating by more than ε from the target output. The ε value determines the tube's width, with a smaller ε indicating less error tolerance, which affects both the number of support vectors and the sparsity of the solution. SVR considers the ε-insensitive loss function as outlined in equation (3) to achieve its goal.

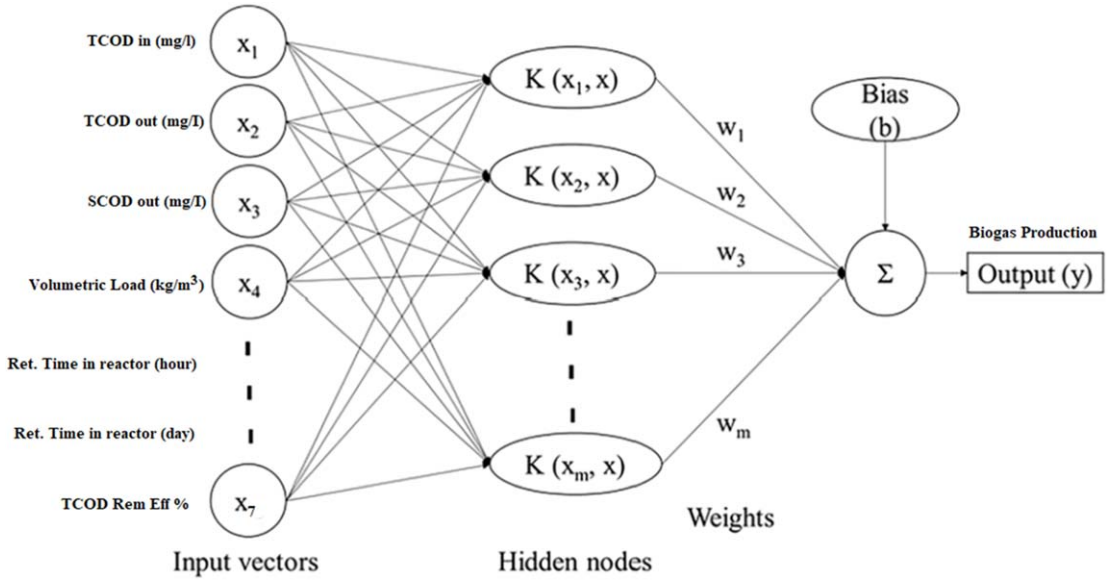

Figure 1. Architecture of the SVR.

$$L_\varepsilon(y, \quad f(x, w)) = \begin{cases} 0 & |y - f(x, w) \leqslant \varepsilon| \\ |y - f(x, w)| - \varepsilon & otherwise \end{cases} \tag{3}$$

SVM uses kernel functions to transform data points from their original feature space into a higher-dimensional space, facilitating linear separation in the new space [31]. The RBF shown in equation (4) is selected as the kernel function because it is the most commonly used kernel and performs well on complex nonlinear data. The RBF kernel computes the exponential of the negative squared Euclidean distance between feature vectors x and x', scaled by a parameter σ. x and x' represent the input feature vectors.

$$k_{RBF}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{4}$$

The gamma value is important for the RBF kernel. A low gamma value provides a broader generalisation but reduces the complexity of the model. A high gamma value, on the other hand, leads to more detailed learning but increases the risk of overfitting. In this study, it is set to 0.7543. For this study, an SVR model is employed, where careful parameter selection is essential to optimise accuracy and generalisation. The input layer comprises seven vectors: TCOD in, TCOD out, SCOD out, volumetric load, retention time in the reactor (hour), retention time in the reactor (day), and TCOD removal efficiency (%). The model features one hidden node and a single output node representing biogas production (figure 1). Implemented in Python using scikit-learn, the model aims to minimise the loss function effectively.

Evaluation of statistical metrics

To assess the robustness of the developed SVR model, Cross-validation, Mean Squared Error (MSE) and the Coefficient of Determination (R^2) have been employed. Cross-validation is a model validation technique that tests how the results of a statistical analysis will perform on an independent dataset. In a predictive problem, the model is typically trained on the known datasets and tested on the unknown datasets (validation set). This testing aims to identify overfitting or selection bias [32]. In this study, 10-fold cross-validation was applied, where the dataset was partitioned into 10 equal subsets (folds).

MSE is a widely used predictor to quantify the average squared difference between the observed and predicted values using equation (4). MSE is derived from the square of the Euclidean distance, which always gives a positive value and decreases as the error approaches zero. For SVR similarity approach, MSE predictor is calculated using equation (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \tag{4a}$$

where n is the number of the data points, y_i and y'_i observed and predicts values for the i-th observation, respectively.

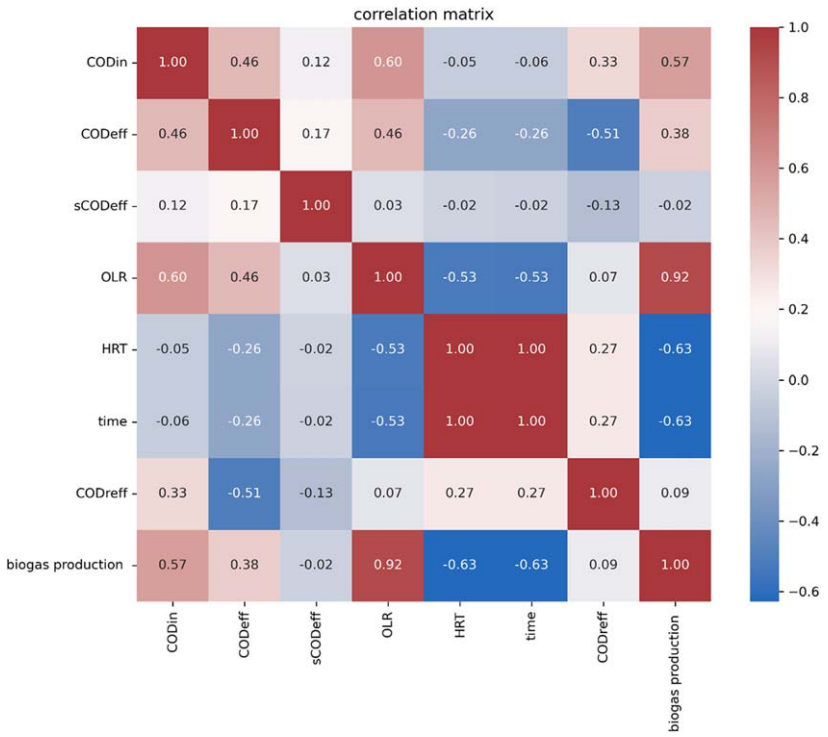

Figure 2. Feature correlation matrix for SVR.

The coefficient of determination (R^2) indicates the extent to which the model accurately replicates observed outcomes, reflecting the proportion of the total variation in the outcomes that is explained by the model. The R^2 is calculated using equation (5):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{5}$$

where n, y_i and y'_i are the number of the data points, observed and predicts values of the i-th observation, respectively like equation (4). In addition, $\overline{y_i}$ it represents the mean of the observed values.

Experimental results

Performance of the SVR model

In this study, an SVR model with RBF kernel is developed for the modelling of biogas production from an anaerobic digester treating fruit processing wastewater. The dataset used encompasses critical operational and performance parameters, highlighting the fundamental dynamics of anaerobic digester performance and its correlation with biogas production. To further examine the relationships between selected inputs and output, a correlation matrix was computed, as presented in figure 2. The analysis revealed that OLR exhibits the strongest positive correlation with biogas production (correlation coefficient = 0.92), indicating its key role as a predictor. In contrast, HRT demonstrates a strong negative correlation (−0.63), suggesting an inverse effect. CODin and CODeff also show moderate positive correlations (0.57 and 0.38, respectively), while sCODeff appears to have negligible influence (−0.02). These results confirm that the five selected features collectively provide a robust basis for modeling, with OLR emerging as the most informative variable in relation to the output.

The performance of the model first comes through in a comparison between actual and predicted values of biogas production, represented in figure 3. The scattered points exhibit a clear trend aligning closely with the perfect fit line (y = x), indicating that the SVR model effectively captures the nonlinear relationships between the input variables and biogas production, demonstrating high predictive accuracy across the entire dataset.

Figure 4 presents a time-series plot of actual and predicted values, demonstrating the model's ability to accurately follow trends and fluctuations in biogas production over time. Performance is demonstrable in both low and high output ranges, attesting to the model's effectiveness under a variety of operational regimes.

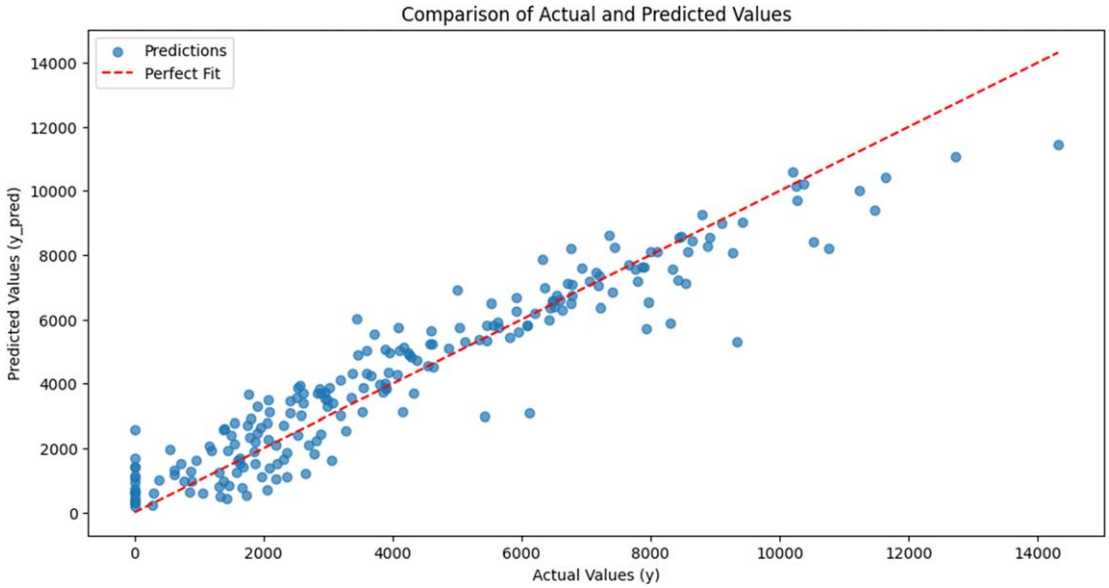

Figure 3. Comparison of actual and predicted biogas production values.

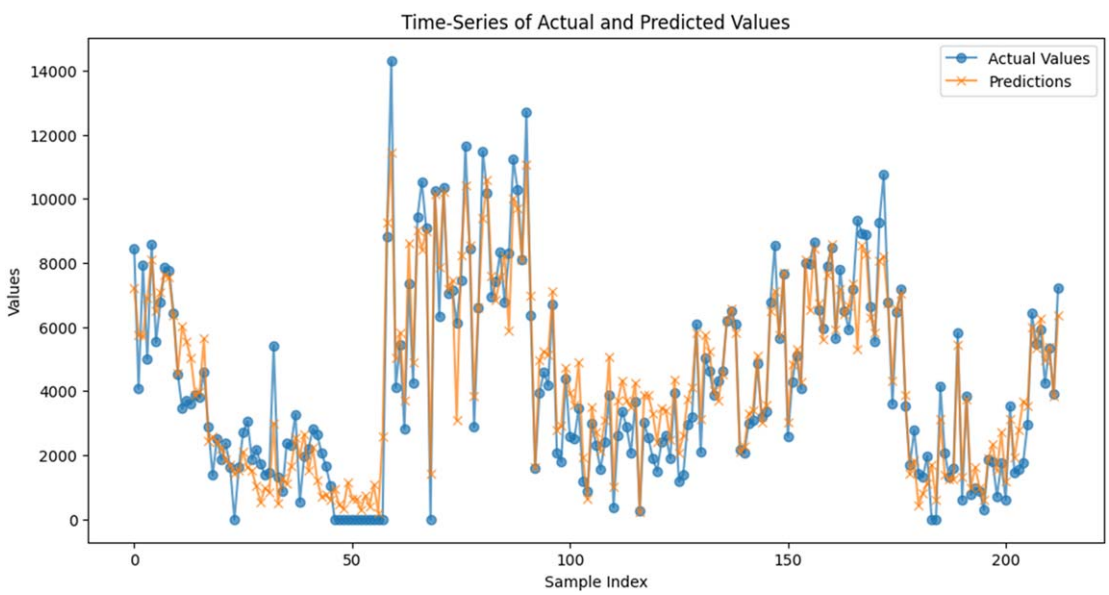

Figure 4. Time-series of actual and predicted biogas production values.

Table 2. Benchmarking MSE performance of regression models via 10-Fold Cross-Validation.

Model	Mean	Standard deviation	Min(Fold-wise)	Max(Fold-wise)
SVR	0.0047	0.0017	0.0011	0.0092
Linear	0.0057	0.0028	0.0018	0.0100
Polynomial (2nd Degree)	0.0086	0.0070	0.0015	0.0260
Polynomial (3rd Degree)	0.0691	0.1424	0.0059	0.4895

Robustness and generalisability of SVR model

The comparative performance of different regression models using 10-fold cross-validation is presented in table 2 (MSE) and table 3 (R^2). These measures provide a sense of the stability and generalisability of the SVR model compared to Linear and Polynomial Regression approaches. The benchmark values in tables 2 and 3 were

Table 3. Benchmarking R^2 performance of regression models via 10-Fold cross-validation.

Model	Mean	Standard deviation	Min(Fold-wise)	Max(Fold-wise)
SVR	0.8983	0.0382	0.8029	0.9248
Linear	0.8707	0.0506	0.7753	0.9434
Polynomial (2nd Degree)	0.7874	0.1881	0.2928	0.9623
Polynomial (3rd Degree)	−0.8693	3.8921	−12.3063	0.8521

established by evaluating the consistency and predictive accuracy of multiple regression models using standard cross-validation metrics. MSE and R^2 were used to quantify both model bias and variance.

The SVR model is the best among all the models regarding MSE, having the lowest error rate (0.0047 mean MSE) and the smallest variation (0.0017 standard deviation), as shown in table 2. This indicates that SVR provides accurate predictions with minimal variation across different folds, reflecting its stability. In table 3, SVR achieves the highest average R^2 value (0.8983), demonstrating that it can account for most of the data variance. Moreover, a low standard deviation of 0.0382 indicates strong generalisation performance, with the model maintaining high predictive quality across various training and validation sets.

Linear Regression (LR) performs really well, with a competitive R^2 score (0.8707), but a less favourable MSE (0.0057) than SVR. This suggests that linear relationships capture a lot of the structure of the data, yet SVR pushes accuracy even further. PR (2nd Degree) demonstrates more volatility in performance with its larger standard deviation both for MSE (0.0070) and R^2 (0.1881). This suggests that the model overfits on some of the folds and underfits on others, reducing the reliability.

PR (3rd Degree) is the least effective, exhibiting an unstable and negative mean R^2 (−0.8693) alongside an exceedingly high standard deviation (3.8921). The significant variation in the fold-wise scores reinforces the existence of extreme overfitting, rendering the model highly unreliable for generalisation. The results from both MSE and R^2 metrics clearly indicate that SVR offers the best balance of accuracy, stability, and generalizability. Its low error, high explanatory power, and minimal performance variation between folds demonstrate its robustness. Unlike polynomial regression, which is highly unstable due to overfitting, SVR successfully captures nonlinear patterns without exhibiting high variance, making it the most dependable regression technique for this data.

Refinement of SVR model

Hyperparameter optimisation plays a crucial role in enhancing machine learning models by identifying the optimal combination of parameters to maximise predictive performance. To refine the SVR model, hyperparameter tuning was conducted using RandomizedSearchCV, which efficiently explores the search space by evaluating a randomly selected subset of parameter combinations. Unlike GridSearchCV, which exhaustively tests all possible configurations, RandomizedSearchCV reduces computational costs while maintaining a high likelihood of identifying near-optimal hyperparameters. By carefully adjusting the number of iterations, it is possible to strike a balance between optimising model performance and ensuring computational efficiency.

To achieve a well-tuned model, a total of 100 different hyperparameter configurations were evaluated using 10-fold cross-validation, leading to 1.000 model fits. The optimisation process determined the best values for γ and C as 0.7543 and 0.3556, respectively, which resulted in a minimum mean squared error of 0.0011. The stability of the cross-validation mean squared error across folds confirms that the refined model maintains high accuracy while avoiding both overfitting and underfitting. This consistency across validation sets indicates that the optimised SVR model generalises well to unseen data, reinforcing its robustness and predictive reliability. In order to further cross-check the validity of the model, residual analysis was performed, as evident in figure 5. The residuals exhibit a uniform distribution centred at zero with no systematic distortion and recognisable patterns. This observation indicates that the model effectively captures the underlying patterns within the dataset with a good margin of accuracy. Furthermore, the absence of heteroscedasticity among the residuals further guarantees the stability of the model and that prediction errors are constant at different levels of the response variable. Taken together, the findings conclude that the optimised SVR model is not only reliable and capable of making predictions with a good margin of accuracy at different working conditions.

Feature contribution analysis using SHAP

To enhance interpretability of the SVR model and provide a deeper understanding of how each input feature influences biogas prediction, SHAP analysis was employed. This post hoc method decomposes the SVR output into additive feature contributions, enabling a detailed assessment of input importance over the entire dataset.

The SHAP results indicate that TCOD Outflow (mg/l) and Volumetric Load (kg/m^3) are the most impactful features, with mean SHAP contributions of approximately +0.0022 and +0.0014, respectively. These features

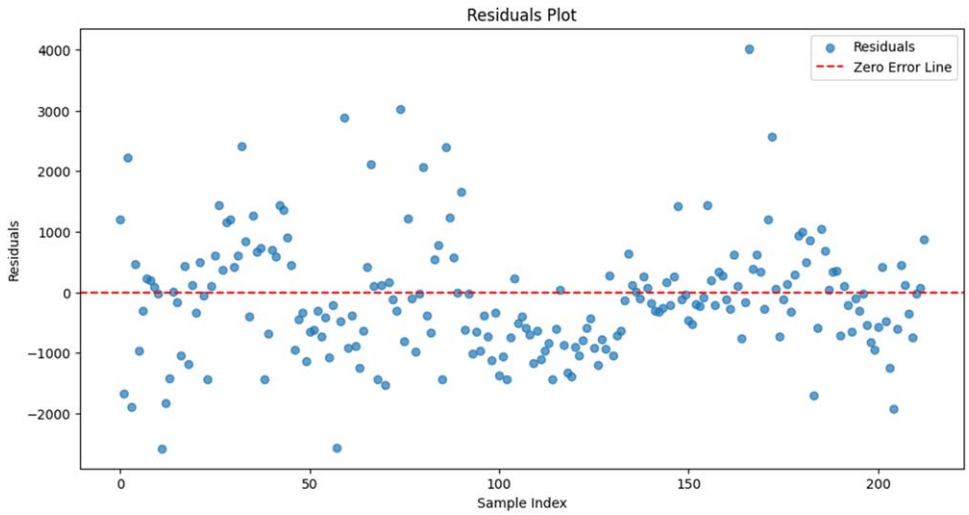

Figure 5. Residuals plot for biogas production predictions.

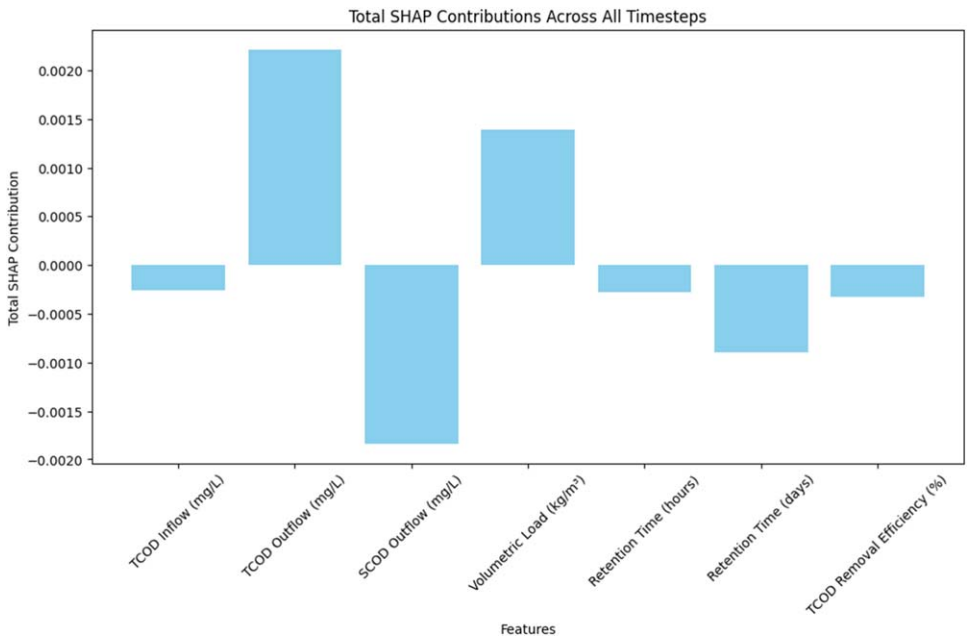

Figure 6. Total SHAP contributions of input features to SVR predictions.

consistently exhibit strong positive influence on biogas prediction, aligning with their established role in anaerobic digestion processes. In contrast, SCOD Outflow (mg/l) shows a negative mean SHAP value around −0.0019, suggesting an inverse or non-contributory relationship with the model output. Retention Time (days) also yields a moderate negative contribution (−0.0012), while other features such as TCOD Inflow and TCOD Removal Efficiency have marginal impact (near zero mean SHAP values).

These findings confirm that predictive relevance is feature and context dependent. As shown in figure 6, TCOD Outflow and Volumetric Load dominate the model's output attribution, underscoring their role as key drivers in SVR-based biogas prediction models.

Comparison of developed models with literature studies

Recent literature underscores the effectiveness of various machine learning (ML) models—such as artificial neural networks (ANN), support vector machines (SVM), adaptive neuro-fuzzy inference systems (ANFIS), and their hybrid forms—in predicting biogas production, methane emissions, and other bio-process parameters, as summarised in table 4. For instance, Abu Qdais *et al* (2010) reported a high predictive accuracy ($R^2 = 0.8703$) using an ANN-GA hybrid model, based on input variables such as temperature, solids content, and pH.

Table 4. Literature-based performance comparison of ML techniques in anaerobic digestion prediction.

Reference	Data size	Model type	Output	R^2	Error rate
(Abu Qdais *et al* 2010)	177	ANN-GA	Biogas production	~0.870	MSE: 0.006
(Xu *et al* 2014)	50	ANN	Methane emission	0.912–0.976	—
[17]	9327	ANN ANFIS SVM	Methane emission	0.680–0.900	MSE: 0.04–0.05 RMSE: 10.31–13.13
(Zaied *et al* 2020)	15	ANN-PSO	Biogas production	0.977–0.998	MSE 0.016–0.209
(Asadi and McPhedran, 2021)	15	ANN-ANFIS	Biogas production	~0.810	RMSE 0.95
(Alejo *et al* 2018)	37	SVMANN	Protein degradation	0.875–0.898	MSE 0.095–0.122
(Olatunji *et al* 2024)	14–18	SVR	Biogas production	~0.900	RMSE 0.0842
(Farzin, *et al* 2024)	297	SVR-GASVR-PSOANN-GAANN-PSO	Biogas production	0.645–0.773	MSE 0.200–0.265RMSE 0.477 - 0.515
[24]	45	SVR	Biogas production	~0.738	RMSE 5.05
Our study	300	SVR	Biogas production	0.802–0.924	MSE 0.001–0.009

Similarly, Xu *et al* (2014) achieved R^2 values ranging from 0.912 to 0.976 for methane emission estimation using ANN architectures applied to various biomass inputs. In larger datasets, such as the one analysed by Mehrdad *et al* (2021), both ANFIS and SVM models demonstrated robust performance, with R^2 values between 0.70 and 0.90, and RMSE values in the range of 11–13, indicating their reliability in modelling landfill methane emissions.

More recent studies have continued to investigate hybrid and optimised ML frameworks. Zaied *et al* (2020) and Asadi and McPhedran (2021) employed ANN-PSO and ANN-ANFIS models, respectively, reporting exceptionally high prediction accuracies (R^2 up to 0.998) in biogas yield prediction. Likewise, support vector regression (SVR) and its optimised variants—such as SVR-GA and SVR-PSO—have shown consistent performance in studies by Olatunji *et al* (2024) and Farzin *et al* (2024), with R^2 values ranging from 0.645 to 0.773 and RMSE values as low as 0.0842. Additionally, Qi *et al* (2022) demonstrated the applicability of SVR in modelling anaerobic baffled reactor (ABR) performance, achieving a moderate R^2 value of 0.738. Collectively, these studies emphasise that model selection, hyperparameter optimisation, and hybridisation strategies significantly influence predictive performance—particularly when aligned with the characteristics of the input variables and the scale of the data.

Limitations of the study

Although the proposed SVR model exhibits high predictive performance, several limitations must be considered to properly interpret its applicability and scope. The model was trained using data from a single full-scale UASB reactor treating wastewater from a fruit processing facility. While this dataset offers rich variability due to seasonal fluctuations in influent composition, it also reflects the specific operational characteristics of one industrial site. Therefore, generalising the findings to other reactor types or wastewater sources—such as municipal or dairy effluents—requires further validation across multiple case studies

Another limitation stems from the number of input variables included in the modelling process. The study utilised five key parameters (CODin, CODeff, sCODeff, OLR, and HRT) based on their known relevance in biogas production and consistent availability in plant records. However, other influential factors—such as temperature, pH, alkalinity, total solids, and microbial community composition—were excluded due to incomplete or irregular data collection. This exclusion may limit the model's ability to capture certain biological or environmental dynamics that affect methane yield.

The learning architecture used in this work is offline and static in nature, relying on historical datasets. While effective for training and validation, such models may not respond optimally to sudden changes in system conditions unless periodically retrained. For deployment in real-time control systems, adaptive or online learning strategies should be considered to enhance responsiveness.

Finally, although SVR is highly effective in learning nonlinear relationships, it functions as a black-box model with limited interpretability. Unlike mechanistic models such as ADM1, it does not provide insight into causal pathways or internal process states. This restricts its utility in diagnostic or explanatory scenarios where understanding of system behaviour is required. Additionally, no uncertainty quantification or input sensitivity analysis was performed, which could have strengthened the robustness evaluation of the model under varying operational regimes.

Conclusion

This study evaluated the predictive capability of a Support Vector Regression (SVR) model with a Radial Basis Function (RBF) kernel for estimating biogas production in a full-scale Upflow Anaerobic Sludge Blanket (UASB) reactor treating industrial wastewater from a fruit processing facility. Using five key operational parameters—namely influent and effluent COD concentrations, soluble COD, volumetric organic loading rate, and hydraulic retention time the SVR model achieved high accuracy, with an average coefficient of determination (R^2) of 0.8983 and a mean squared error (MSE) of 0.0047 under 10-fold cross-validation.

Comparative analysis showed that SVR consistently outperformed conventional regression approaches, particularly third-degree Polynomial Regression, which exhibited overfitting and poor generalisation. The robustness of the SVR model under conditions of variable influent composition further underlines its suitability for real-world industrial applications. Seasonal fluctuations in wastewater characteristics, often a challenge for modelling efforts, were effectively captured without degradation in predictive performance. Hyperparameter tuning via RandomizedSearchCV was instrumental in enhancing model generalisation while maintaining computational efficiency.

In addition to its predictive strength, the model's interpretability was enhanced using SHAP analysis, which quantified the contribution of each input variable to the model's output. TCOD Outflow and Volumetric Load emerged as the most influential predictors, while features such as SCOD Outflow and Retention Time had lower or even negative contributions. These insights not only align with established process knowledge but also

demonstrate the model's transparency and reliability. The integration of SHAP analysis supports explainable AI practices, enabling process engineers to make informed decisions based on both model outputs and the underlying feature dynamics.

Overall, the SVR model presents a reliable and interpretable solution for biogas yield prediction and process optimisation in anaerobic digestion systems. Future studies may explore the inclusion of additional process parameters such as temperature, pH, and alkalinity, as well as the integration of real-time sensor data streams. Furthermore, hybrid approaches that couple machine learning models with mechanistic frameworks like ADM1 could offer a powerful synergy between predictive accuracy and process-level interpretability, paving the way toward intelligent and adaptive biogas production systems.

Statement of conflict of interest

The authors declare there is no known conflict of interest.

Funding

This research has not received funding from any organization.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

M. Erdem Isenkul https://orcid.org/0000-0003-0856-2174
Sevgi Güneş-Durak https://orcid.org/0000-0003-4273-7417
Yasemin Poyraz Kocak https://orcid.org/0000-0002-1502-7260
İnci Pir https://orcid.org/0000-0002-0540-5387
Mertol Tüfekci https://orcid.org/0000-0002-5530-1471
Güler Türkoğlu Demirkol https://orcid.org/0000-0001-8222-5497
Selçuk Sevgen https://orcid.org/0000-0003-1443-1779
Aslı Seyhan Çığgın https://orcid.org/0000-0002-3290-7112
Neşe Tüfekci https://orcid.org/0000-0002-7373-4131

References

[1] Gupta R, Zhang L, Hou J, Zhang Z, Liu H, You S, Sik Ok Y and Li W 2023 Review of explainable machine learning for anaerobic digestion *Bioresour. Technol.* **369** 128468
[2] Garcia-Tirado R, Fernandez-Crespo E, Font X, Vicent T, Peralta J, Trifi D, Martinez-Cuenca R and Chiva S 2024 Long-term performance and activity study of a two-stage anaerobic EGSB reactors system treating complex and toxic industrial wastewater *Water Environ. Res.* **96** e11109
[3] Güneş Durak S, Acarer S and Türkoğlu Demi˙rkol G 2023 Treatment of citrus juice process wastewater with UASB and biogas production *Environ. Res. Technol.* **6** 68–77
[4] Lamoh M M, Joy E J, Rashid M, Islam S and Hashmi S 2020 Biogas production optimization from palm oil mill effluent: experiments with anaerobic reactor *Int. J. Integr. Eng.* **12** 261–70
[5] Sun H, Yu N, Mou A, Yang X and Liu Y 2022 Effluent recirculation weakens the hydrolysis of high-solid content feeds in upflow anaerobic sludge blanket reactors *J. Environ. Chem. Eng.* **10**
[6] Neves A, Roseiro L B, Ramalho L, Eusébio A and Marques I P 2020 Hybrid anaerobic reactor: brewery wastewater and piggery effluent valorisation *Wastes: Solutions, Treatments and Opportunities III - Selected papers from the 5th International Conference Wastes: Solutions, Treatments and Opportunities, 2019*
[7] Ciggin A S 2016 Anaerobic co-digestion of sewage sludge with switchgrass: experimental and kinetic evaluation *Energy Sources, Part A Recover Util. Environ. Eff.* **38** 15–21
[8] Kelleher B P, Leahy J J, Henihan A M, O'Dwyer T F, Sutton D and Leahy M J 2002 Advances in poultry litter disposal technology—a review *Bioresour. Technol.* **83** 27–36
[9] Haugen F, Bakke R and Lie B 2013 Adapting dynamic mathematical models to a pilot anaerobic digestion reactor *Model. Identif. Control* **34** 35–54
[10] Andrade Cruz I, Chuenchart W, Long F, Surendra K C, Renata Santos Andrade L, Bilal M, Liu H, Tavares Figueiredo R, Khanal S K and Fernando Romanholo Ferreira L 2022 Application of machine learning in anaerobic digestion: perspectives and challenges *Bioresour. Technol.* **345** 126433
[11] Barbur V A, Montgomery D C and Peck E A 1994 introduction to linear regression analysis *Stat* **43** 339–41
[12] Roustaei N 2024 Application and interpretation of linear-regression analysis *Med hypothesis, Discov Innov Ophthalmol J* **13** 151–9

[13] Zhou W, Yan Z and Zhang L 2024 A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction *Sci. Rep.* **14** 5905
[14] Gabauer D, Gupta R, Marfatia H A and Miller S M 2024 Estimating U.S. housing price network connectedness: evidence from dynamic elastic net, lasso, and ridge vector autoregressive models *Int. Rev. Econ. Financ.* **89** 349–62
[15] Callens A, Morichon D, Abadie S, Delpey M and Liquet B 2020 Using random forest and gradient boosting trees to improve wave forecast at a specific location *Appl. Ocean Res.* **104** 102339
[16] Yu S, Ren Y, Wu X, Guo P and Li Y 2024 Dynamic pruning-based bayesian support vector regression for reliability analysis *Reliab. Eng. Syst. Saf.* **244** 109922
[17] Mehrdad S M, Abbasi M, Yeganeh B and Kamalan H 2021 Prediction of methane emission from landfills using machine learning models *Environ. Prog. Sustain. Energy* **40** e13629
[18] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L and Lopez A 2020 A comprehensive survey on support vector machine classification: applications, challenges and trends *Neurocomputing* **408** 189–215
[19] Awad M and Khanna R 2015 *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* 1st (Apress Open, New York)
[20] Zhai S, Chen K, Yang L, Li Z, Yu T, Chen L and Zhu H 2024 Applying machine learning to anaerobic fermentation of waste sludge using two targeted modeling strategies *Sci. Total Environ.* **916** 170232
[21] De Gregorio L, Callegari M, Mazzoli P, Bagli S, Broccoli D, Pistocchi A and Notarnicola C 2018 Operational river discharge forecasting with support vector regression technique applied to alpine catchments: results, advantages, limits and lesson learned *Water Resour. Manag.* **32** 229–42
[22] Akram M T, Aftab R A, Ansari K B, Arman I, Hakeem M A, Zaidi S and Danish M 2024 Innovative approach to characterize cheese whey anaerobic digestion using combined mechanistic and machine learning models *Bioenergy Res* **17** 2474–86
[23] Liu L and Lei Y 2018 An accurate ecological footprint analysis and prediction for Beijing based on SVM model *Ecol. Inform.* **44** 33–42
[24] Qi Z, Wang Z, Chen M and Xiong D 2022 Pilot-scale anaerobic treatment of printing and dyeing wastewater and performance prediction based on support vector regression *Fermentation* **8** 99
[25] Granata F, Papirio S, Esposito G, Gargano R and de Marinis G 2017 Machine learning algorithms for the forecasting of wastewater quality indicators *Water (Switzerland)* **9** 105
[26] Ke B, Nguyen H, Bui X N, Bui H B, Choi Y, Zhou J, Moayedi H, Costache R and Nguyen-Trang T 2021 Predicting the sorption efficiency of heavy metal based on the biochar characteristics, metal sources, and environmental conditions using various novel hybrid machine learning models *Chemosphere* **276** 130204
[27] Parajuli A, Khadka A, Sapkota L and Ghimire A 2022 ffect of hydraulic retention time and organic-loading rate on two-staged, semi-continuous mesophilic anaerobic digestion of food waste during start-up *Ferment 2022* **8** 620 8:620
[28] APHA, AWWA, WEF 2012 *Standard Methods For Examination Of Water And Wastewater.* 22nd ed (American Public Health Association)
[29] Vapnik V N 2000 *The Nature of Statistical Learning Theory* 2nd ed (Springer LLC)
[30] Vapnik V N and Chervonenkis A Y 1974 On the method of ordered risk minimization *Avtomat i Telemekh* **35** 1226–35 https://m.mathnet.ru/links/30351cf11347f2c5f72c4bcc52cf7b1d/at8452.pdf
[31] Ma T and Niu D 2016 Icing forecasting of high voltage transmission line using weighted least square support vector machine with fireworks algorithm for feature selection *Appl. Sci.* **6**
[32] Cawley G C and Talbot N L C 2010 On over-fitting in model selection and subsequent selection bias in performance evaluation *J. Mach. Learn. Res.* **11** 2079–107 https://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf