fermentation

MDPI

Article

Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process

Senem Önen Cinar *, Samet Cinar and Kerstin Kuchta

Circular Resource Engineering, Economy and Management, Hamburg University of Technology, Blohmstr. 15, 21079 Hamburg, Germany; samet.cinar@tuhh.de (S.C.); kuchta@tuhh.de (K.K.)
* Correspondence: senem.oenen@tuhh.de

Abstract: Process optimization is no longer an option for processes, but an obligation to survive in the market in any industry. This argument also applies to anaerobic digestion in biogas plants. The contribution of biogas plants to renewable energy can be increased through more productive systems with less waste, which brings the common goal of minimizing costs and maximizing yields in processes. With the help of data science and predictive analytics, it is possible to take conventional process optimization and operational excellence methods, such as statistical process control and Six Sigma, to the next level. The more advanced the process optimization aspect, the more transparent and responsive the systems. In this study, seven different machine learning algorithms—linear regression, logistic regression, K-NN, decision trees, random forest, support vector machine (SVM) and XGBoost—were compared with laboratory results to define and predict the possible impacts of wide range temperature fluctuations on process stability. SVM provided the best accuracy with 0.93 according to the metric precision of the models calculated using the confusion matrix.

Keywords: temperature management; anaerobic digestion; process optimization; machine learning

Citation: Cinar, S.Ö.; Cinar, S.; Kuchta, K. Machine Learning Algorithms for Temperature Management in the Anaerobic Digestion Process. *Fermentation* **2022**, *8*, 65. https://doi.org/10.3390/fermentation8020065

Academic Editors: Steven Wainaina and Mukesh Kumar Awasthi

Received: 31 December 2021
Accepted: 28 January 2022
Published: 30 January 2022

1. Introduction

Germany has been converting its energy sources from coal, oil and nuclear to renewable energy sources for more than 40 years. The planned reduction in greenhouse gas with Germany's 'Climate Action Plan 2050′ based on 1990 is as follows: 40% by 2020, 55% by 2030, 80–95% by 2050 [1].

The German government aims to increase the contribution of renewable energy to 65% by 2030. For this reason, the need for new, low-CO_2, controllable capacities to cover the residual load is foreseeable [2]. Biogas plants have an enormous role in this strategy. During the anaerobic digestion (AD) process, biomass is converted into biogas through a series of complex biochemical reactions. In a Combined Heat and Power (CHP) Plant unit, produced gas is processed to produce heat and electricity. Usually, there are a storage area for biomass, pre-treatment, digester, biogas processing units and digestate processing units in a biogas plant. The anaerobic digestion process has four sequential stages: hydrolysis, acidogenesis, acetogenesis and methanogenesis. Since there are different kinds of microorganisms taking part in each stage of the anaerobic digestion process, supplying a suitable environment for a sustainable process is complex and makes it challenging monitor and control the process [3]. Any kind of anomaly from one of these stages can cause an unstable process. With the help of monitoring, it is possible to track every possible failure in a system and obtain an early response to these risk components. In general, monitoring of the processes is mandatory to obtain an overview of the general process and be warned of possible anomalies before they occur. Standard monitoring includes monitoring of daily substrate feeding amount, daily gas production amount, daily temperature management in the reactor, pH measurement twice a week, biogas quality (1–2 times per week), VOA/TIC (Volatile Organic Acids/Total Inorganic Carbon) twice per month, VFA (Volatile Fatty Acids)

1–2 times per month. However, due to the various biological interactions during anaerobic digestion, it is not possible to run standard monitoring for all systems. Nevertheless, it is essential to define relevant process parameters such as temperature, pH, pressure, the mass of input and organic load rate [4,5]. Deviating feed rates and intervals, temperature changes, ammonia inhibition, hydrogen sulfide inhibition, and other inhibitory substrates in the feedstocks are potential process disturbances [4]. It is also possible to create a system with online monitoring and threshold control. With the help of online monitoring and predictive statistics, Programmable Logic Controllers (PLC) could enable us to have automated systems with defined thresholds [6]. As stated in the 'Germany 2020 Energy Policy Review' and 'Flexibilization of Biogas plants', Germany's primary energy strategy is to invest in renewable energy such as biogas [2]. Enhancing the operational efficiency and contribution of biogas plants to renewable energy can be achieved by reducing operating costs, flexible systems and increasing the quality.

Anaerobic fermentation refers to the decomposition of organic material by microorganisms in an oxygen-free environment. The biogas produced by this process consists mainly of methane (CH_4) and carbon dioxide (CO_2) [7]. There have already been many studies conducted about process optimization in anaerobic digestion and biogas plants, based on different parameters on both a laboratory and an industrial scale. The effect of temperature on AD has also been one of the main focuses in this area. In the study performed by Tian et al. (2018), temperature changes from 9 to 55 °C were analyzed in relation to the behavior of the various microbial communities and metabolic pathways involved in AD. The study showed an increase for metabolic activities when temperature increased from 15 to 35 °C [8]. The digester performance and its relation to operating parameters such as temperature were investigated by Westerholm et al. (2015). During the 320 days of operation, operation temperatures of 37 and 42 °C were analyzed in this research [9]. K.J. Chae et al. (2008) observed different yields of produced methane under different conditions regarding temperature and the amount of feed [10]. In the scope of the study, the next-generation sequencing (NGS)-based metagenomics approach was the monitoring technic for mesophilic (37 °C) and thermophilic (55 °C) environments. The aim was also to determine the operational parameters of the anaerobic degradation process [11]. Member and Sallis studied, in 2018, the effect of temperature on anaerobic digestion in microalgae [12]. The increasing trend of biogas plants makes it necessary to work on process optimization in biogas plants. That is why the effects and prevention of the overheating of fermenters during summertime were studied by Bavutti et al. (2014) [13]. In order to define process stability, temperature fluctuations in an industrial-scale digester from two different wastewater treatment plants were analyzed. For different conditions, chemical oxygen demand balances were the scope of Hubert et al. (2019) [14]. In another study performed by Terradas et al. (2014), possible heat transfer between a biogas digester and the soil surrounding it was analyzed through a noncomplex setup, with input data excluding most of the operational parameters such as heating, stirring and the insulating digester, which were buried in the soil [15]. Due to several reasons, such as inadequate heating systems, overheating problems, insufficient mixing, insufficient insulation and extreme weather conditions, temperature management in biogas plants can be challenging. On the other hand, the flexibility of temperature in the reactor can supply efficient usage of heat and innovative control of processes via implementing different scenarios, e.g., different operation temperatures in different season [13]. The laboratory analyses in this study aimed to examine the impact of temperature regime changes on process efficiency with different feeding rates.

The increase in interest in biogas plants makes it necessary to work on process optimization. Implementation of machine learning algorithms can help in the better understanding of temperature impact on the anaerobic digestion process. Traditional statistical methods can also be understood as traditional process optimization approaches such as Statistical Process Control (SPC) and Six Sigma, where we need statistics to improve quality and optimize processes. For this reason, Lean Six Sigma approaches have been used in various

process industries [16]. There have already been some statistical methods and Artificial Intelligence (AI) studies conducted in the area of living (biological) systems. In order to be able to take advantage of predictive analytics, it is obligatory to get a sufficient number of data to work with. In dynamic systems in particular, such as biogas plants, where there are not only internal parameters but also external parameters affecting the process, real-time monitoring has enormous importance. Online monitoring comes with the requirement of parameter definition being monitored, tracked and analyzed [17].

The usage of online monitoring systems is not very common among biological systems, including anaerobic digestion. Implementation of these methods could surely bring better data gathering for statistical process optimization [18]. Analyzing and obtaining valuable outputs from a considerable number of gathered data could be possible with only mathematical and statistical methods. These analyses also allow for predictive analytics and continuous improvement, which allow not only flexibility in energy production but also a robust system [19,20]. Standardizing the requirements of process optimization is not possible in dynamic systems. Living organisms cannot be defined as stable systems in terms of their parameters and traceability, so it is not easy to track them with conventional monitoring methods. A variety of characters and uncontrollable environmental conditions bring indifferent demands on feeding and maintaining in AD. This dynamic behavior can be monitored and measured by modern methods, and interpreted accurately [21]. There have also been machine learning applications in the natural sciences that enable intelligent decision-making systems and artificial and computational intelligence in the data science. Daily time series data in a wastewater treatment plant were used as input for developing a support-vector machine model by Manu and Thalla (2017). It was observed that it is possible to define the relationship between dependent and independent variables with the help of machine learning algorithms [22,23]. Machine learning models give different approaches to predictive analytics, such as regression and classification models. As Wang and Long mentioned in their research in December 2019, recently developed algorithms such as Artificial Neural Network (ANN), SVM, random forest, Logistic Regression Multiclass (GLMNET) were used as both regression and classification models in their study to define sufficient parameters and use prediction about methane production in anaerobic digestion [24]. Nourani et al. (2018) aimed to make a prediction of the performance from a wastewater treatment plant with different models such as feed-forward neural network, adaptive neuro-fuzzy inference system, support vector machine and a multilinear regression were developed. Another study compares different models (Gompertz, machine learning and hybrid) regarding the feasibility of machine learning usage or process optimization in biogas production [25,26]. Last but not least, there have been studies related to predictive maintenance with more modern algorithms and approaches. In 2019, different machine learning models such as logistic regression, support vector machine, random forest, extreme gradient boosting (XGBoost), and k-nearest neighbor regression were compared with a dataset provided by two major Chinese biogas plants on a daily basis. The aim of this study was to develop a user interface with machine learning to improve the productivity in an industrial scale biogas plant. This approach can also be considered as a concrete and industrial usage of predictive analytics [27].

In this study, different machine learning algorithms were compared to make predictions of the daily methane generation volume produced by a laboratory setup (under continuous operation), to be able to define the optimum temperature with other operational parameters in anaerobic digestion.

With the help of predictive analytics, it is possible to answer the abovementioned questions. Lab-scale data were used for the modelling in this study, to integrate machine learning algorithms into biogas plant operation for efficient monitoring.

2. Materials and Methods

2.1. Laboratory Set-Up

In this study, Continuously Stirred Tank Reactors (CSTRs), as represented in Figure 1, were used for performing biogas production tests. Substrates and discharges of the reactors were stored in a fridge under the reactors. Feeding and discharging of reactors were performed via programmable pumps, which were attached to the system automatically. The biogas volume and biogas content data were recorded in a 20-min period through a methane sensor and MilliGascounter. Four identical reactors were used to simulate different scenarios for biogas production. The reactors shown in Figure 2 were equipped with an agitator, a heating mat including insulation, two pumps for the nutrient solution and the fermentation residue, three temperature sensors, a methane sensor and a gas counter.

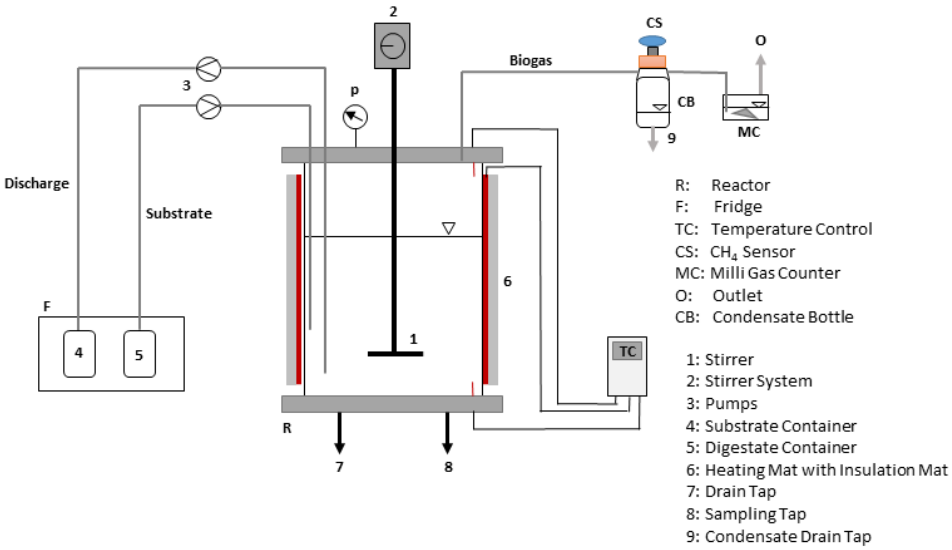

Figure 1. Experimental setup of the continuous fermentation test with CSTR.

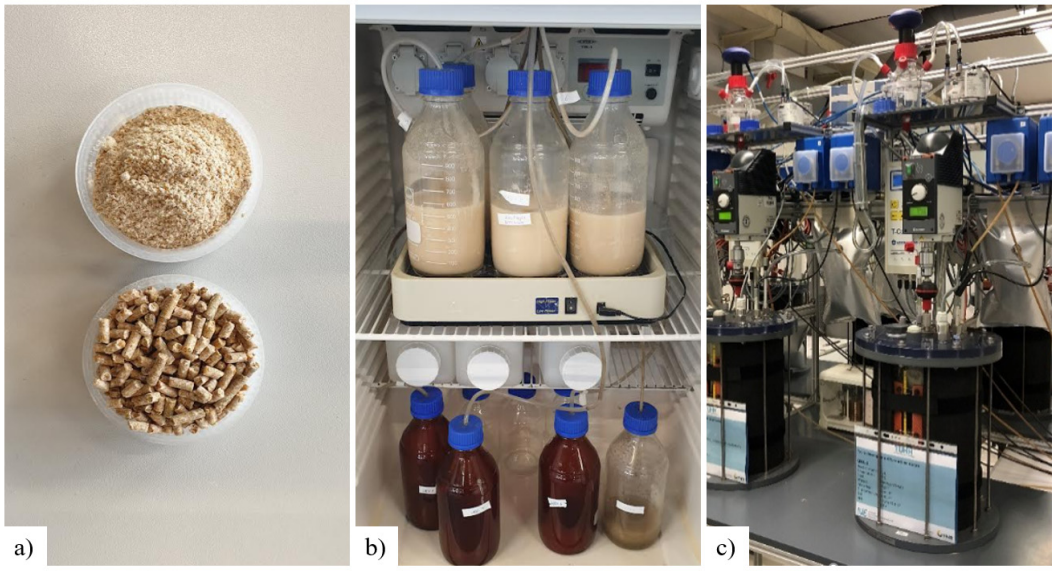

Figure 2. (**a**) Original and grinded pellets; (**b**) feeding and discharging bottles in the fridge; (**c**) CSTR reactors.

These were each filled with 4.5 L of digestate from Digester 1 (first stage of the two-stage anaerobic digestion process) of an industrial biogas plant, and 1 L of distilled water to dilute and supply suitable TS content for the laboratory scale reactors. The

industrial scale digester was under operation at 42 °C and in this digester, the first stage of two-stage anaerobic digestion was conducted. Storage tanks for the nutrient solution and collection tanks for the fermentation residue were stored in a refrigerator integrated into the experimental plant, and were connected from there by pipes to the pumps. In order to avoid blockages in the pipes, the digestate from the biogas plant had to be filtered before feeding into the reactors. Since the digestate had an exceptionally high proportion of corn silage at the time of removal, the volume was reduced by about half during filtering (filter hole size: 1 mm, stainless steel filter). The reactors were initially operated at 42 °C. The RZR 2052 control stirrers from Heidolph were set to 55 rpm.

Pellets (animal feed material) was used as substrate to supply the continuous feeding of reactors. The main components of pellets were as follows: crude protein (10.5%), crude oil/fat (4%), crude fiber (2.7%), crude ash (2%), calcium (0.07%), phosphor (0.3%), sodium (0.02%), lysin (0.38%). The nutrient solution (see Figure 3), with which the reactors were fed semi-continuously, was regularly mixed with 2 L of distilled water and 200 g of pellets. To see the effects of temperature changes at two different Organic Loading Rates (OLR), CSTR 3 and CSTR 4 were fed directly with this nutrient solution. At the same time, for CSTR 1 and CSTR 2, it was again diluted 1:1 with distilled water to supply different OLRs.

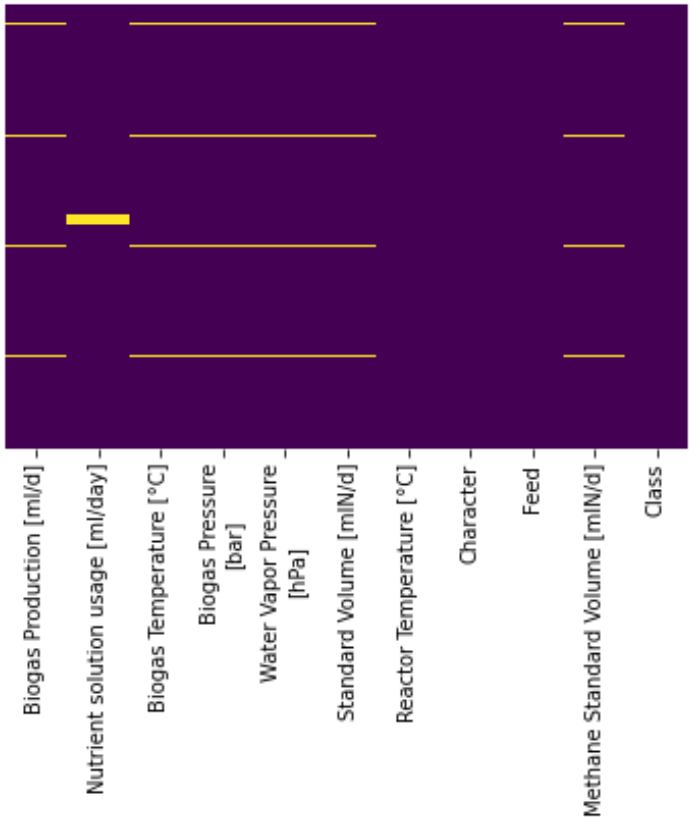

Figure 3. Defining the missing data with the help of a heatmap (*y*-axis shows the data continuity).

After obtaining a stable biogas generation rate from each reactor, temperature changes were performed. The reference temperature was chosen as 42 °C, which is the operation temperature of the industrial scale biogas plant. Based on the mesophilic and thermophilic temperature ranges, the temperature increase/decrease in the reactors was:

- CSTR 1: 42 °C–55 °C–42 °C, OLR of 0.54 g vs. $(L\ d)^{-1}$
- CSTR 2: 42 °C–29 °C–42 °C, OLR of 0.54 g vs. $(L\ d)^{-1}$
- CSTR 3: 42 °C–55 °C–42 °C, OLR of 1.08 g vs. $(L\ d)^{-1}$
- CSTR 4: 42 °C–29 °C–42 °C, OLR of 1.08 g vs. $(L\ d)^{-1}$

In this study, the possible effect of temperature regime changes was evaluated. Therefore, the temperature changes reached 55 °C in the thermophilic temperature range, and decreased the temperature to 29 °C.

In continuous reactors, the response of the anaerobic process to a medium-term change in temperature was studied. The temperature in two reactors with different OLRs was raised by 1 °C per day for 13 days, kept constant for 14 days, and returned to the initial state at the same rate of change. In two other reactors with different OLRs, the temperature was first lowered and raised in the same rhythm. There is no irreversible impact on the process observed in the reactors, and initial biomethane generation amounts were reached at the end of the experiment.

In CSTRs, TS (Total Solids) content, vs. (Volatile Solids) content, pH-Value, VOA/TIC, VFAs, hydrogen carbonate (HCO_3^-), TS, VS, temperature, methane content and biogas volume flow were measured. The analyses were performed according to standards DIN 38-414-S2 (TS), DIN EN 12,879 (S 3a) (VS), DIN 38 404-C5 (pH), DIN 38409-H7-1-2 (HCO_3^-), Nordmann method (VOA/TIC) and DIN 38409-H21 (VFAs).

2.2. Machine Learning

2.2.1. Raw Data Collection, Understanding and Preparation

In this phase of the process, the objective of the project will be defined in detail. Furthermore, for the analysis, data were stored during the data gathering.

The research objective can be determined as studying the flexibility of the process to temperature changes in anaerobic digestion. The aim of the research was to create a model to predict the behavior of anaerobic digestion with the help of data science. For this purpose, a laboratory setup was organized for anaerobic digestion, where temperature fluctuations were realized. Operational data were collected in a 20-min period from four identical CSTR reactors. For each reactor, the output data were collected with the help of methane sensors and MilliGascounters as mentioned above. For data preparation, the various outputs were distributed across multiple spreadsheets that corresponded to the daily production; these datasets were merged into a single data frame for the project.

A basic transformation of the target variable was performed, and the values were divided into three groups: "low" class was defined as the standard methane volume (volume of dry gas in normal state) between 9.91 NmL and 901.82 NmL d^{-1}; "medium" class was defined as a volume between 901.82 NmL d^{-1} and 1707.86 NmL d^{-1}; and "high" class was defined as all values greater than 1707.86 NmL d^{-1}. These classes were encoded with the help of the 'scikit-learn' library, to be able to deploy classification models with the dataset.

2.2.2. Data Interpretation

In the second phase, the aim was to develop a more profound insight into the interpretation of the objective, by describing the collected and verified data. These steps allowed us to conduct the first exploratory analysis before modelling. This phase consisted of four stages; initial data collection, which gave a rough overview of the data at the beginning regarding their characteristics, such as whether they were quantitative or qualitative [27–29]. In the second stage of data interpretation, the data needed to be described to perform the first initial analysis. At this stage, the source, the size, and the types of data were determined. Afterwards, it was possible to launch exploratory data analysis as the third stage in this phase. In the end, this phase was necessary to define the weak points and aspects of the data.

In this study, the libraries 'pandas', 'NumPy', 'Matplotlib' and 'Seaborn' were used through the programming language Python. The first five cells of the original datasets are represented in Table 1.

The dataset has 18 columns. Undesirable and useless column values were dropped from the dataset. The columns with possible impacts on the model were kept for model building.

Table 1. The first five cells of the dataset after dropping the unnecessary columns.

BP [mL d^{-1}]	Nutrient Solution Usage [mL d^{-1}]	BP Temperature [°C]	Biogas Pressure [bar]	Waste Vapor Pressure [hPa]	Standard Volume [NmL d^{-1}]	Reactor Temperature [°C]	Character	Feed	Methane Standard Volume [NmL d^{-1}]	Class
1588.62	67.399	23.197	1.020	28.391	1433.286	42	4	0.5	787.642	Low
2162.71	74.936	23.369	1.019	28.687	1948.434	42	4	0.5	1115.017	Medium
1983.10	74.936	23.632	1.021	29.146	1786.958	42	4	0.5	1055.899	Medium
2001.46	74.936	23.543	1.015	28.988	1793.475	42	4	0.5	1064.220	Medium
1859.20	74.936	21.859	1.011	26.176	1673.445	42	4	0.5	963.959	Medium

BP: Biogas Production.

In order to define the missing data, a simple heatmap (see Figure 3) was used from the 'Seaborn' library. The values with yellow marks are missing from the dataset due to the technical challenges. As seen in the figure, several columns have some missing values due to a possible problem with the sensor or experiment itself. In order to be able to create a statistical model, it was necessary to analyze and defined this missing data. Otherwise, it was possible to affect the accuracy of the model. The heatmap (from Seaborn) usage is one of the ways to define missing data. There is a plethora of methods that can define and handle missing values in a dataset. The process for the missing data will be explained and applied in the next section of the dataset.

2.2.3. Data Preparation

Data preparation in this study was essentially comprised of three parts: encoding the categorical data, data cleansing, and standardization of the variables. In the scope of this study, the aim was to be able to build not only classification models but also regression models. For the classification models, output data were divided into three groups: low, high and medium.

For the first step of data preparation, categorical variables were turned into numerical variables with the help of the library 'sklearn. preprocessing'. This formed the basis for the following steps of data cleansing and feature engineering.

In data cleaning, the goal is to free the data from low quality. With outliers in particular, missing data and inaccurate data strongly influence the quality of the modelling. In order to perform data cleaning successfully, input from the business is needed to clarify which content is correct and which needs to be corrected. The average value of the columns was used to replace the missing values. Thus, it was possible to remove all yellow-marked missing values, as shown in Figure 3 from the dataset.

The last step in data preparation is the standardization of the variables. This involves the process of bringing the data into a uniform format to be able to build a model with the given dataset. These constructed variables are based on the experiences made in the data interpretation phase and are used in the next step, model building.

2.2.4. Model Building

Modelling—i.e., building a statistical equation for the existing data—is the core of the data science process [30]. Usually, methods from the field of machine learning are used in this step to generate a predictive model based on the historical data

After the data preparation step, which includes cleaning and transforming the data, the dataset was divided into a training set and a test set to compare the accuracy of different machine learning algorithms applied to the dataset [31]. Due to their feasibility in biological science and promising predictions, linear regression, logistic regression, k-nearest-neighbors (kNN), support vector machines, decision trees, random forest and XGBoost models were used in this project.

Linear regression is a well-known mathematical method of modelling the relationship between a dependent variable and one or more independent variables. Regression uses

existing (or known) values to predict the required parameters. The linear regression model is an important and useful tool in many statistical analyses for examining the relationship between variables. Standard linear regression assumes that the response variable is scalar. The equation of the linear regression is displayed in Equation (1), where y is the dependent variable, X is the independent variable and b_0 is the constant and b_1 is the coefficient for the independent variable [30,32].

$$y = b_0 + b_1 \times X_1 \quad (1)$$

Logistic regression is the appropriate regression analysis to perform when the dependent variable is binary. This makes logistic regression a classification procedure. As with all regression analyses, logistic regression is a predictive analysis and is used to describe data and explain the relationship between a dependent binary variable and one or more independent variables. Unlike linear regression, in logistic regression, the specific value of the criterion is not predicted. For prediction, a regression equation is implemented in logistic regression. If the regression equation is transferred into a coordinate system, the characteristic curve of the logistic regression can be seen. It can be used to estimate how likely a characteristic expression of the criterion is for a variable with a certain predictor value, and how well the model fits the dataset. The classical linear regression equation (Equation (1)) can be converted to a logistic regression using a sigmoid function [31,33].

$$\text{Sigmoid Function}: \; p = \frac{1}{1 + e^{-y}} \quad (2)$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \times X_1 \quad (3)$$

K-nearest-neighbor is a data classification algorithm that attempts to determine which group a data point belongs to. In is an algorithm that looks at a point in a grid and tries to determine whether a point is in group A or B. The area is arbitrarily determined, but the point is to take a sample of the data. If the majority of the points are in group A, it is likely that the data point in the sample is A and not B, and vice versa. For this project, k value was chosen for tuning the model. To demonstrate a K-nearest-neighbor analysis, consider the task of classifying a new object among a number of known examples [31,32].

Support vector machines (SVM) are supervised machine learning techniques mainly used for binary classification. However, in this project, SVM has also been used for regression analysis as support vector regression (SVR). The training data is plotted in n-dimensional space, and the algorithm tries to draw a boundary with the largest possible distance to the nearest sample. Support vectors are coordinates of individual observation. The support vector machine is a boundary that best separates the classes [31]. With the help of Gridsearch, it was possible to find out the best parameter to tune the model for better prediction. In the scope of this project, the radial basis function (rbf) kernel was used to handle the nonlinearity of the dataset. After launching Gridsearch with the help of Python, the best 'C' and gamma value were identified as 10 and 1, respectively, to make the most accurate predictions.

Decision trees are a kind of supervised machine learning, where data are continuously partitioned according to a certain parameter. Decision trees for regression analysis have already been explained above and illustrated with the help of an application example. Decision tree methodology is a commonly used data mining method for creating classification systems to develop the predictive algorithms for a target variable. This method classifies a population into industry-like segments that create an inverted tree with root, internal, and leaf nodes. The algorithm is a non-parametric structure. If the sample size is large enough, the study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset helps to determine the appropriate tree size needed to achieve the optimal final model [32].

The basis for random forest is formed by many individual decision trees. However, since a single tree consists of several branches, it is possible to produce more satisfactory results with the random forest classification or regression [30,32]. Random forest and decision tree algorithms have been used both for regression and classification models.

All of the algorithms given can be improved by using the boosting method to get better accuracy for the model. Boosting is a general approach to combining different classifiers, in order to achieve an improved overall performance. Gradient boosting is a type of boosting algorithm. It is based on the intuition that the best possible next model in combination with previous models minimizes the overall prediction error. The key idea is to set the target results for this next model to minimize the error. Gradient boosting (XGBOOST) can be used for classification as well as for regression [30]. A general overview of the methods is given in Figure 4.

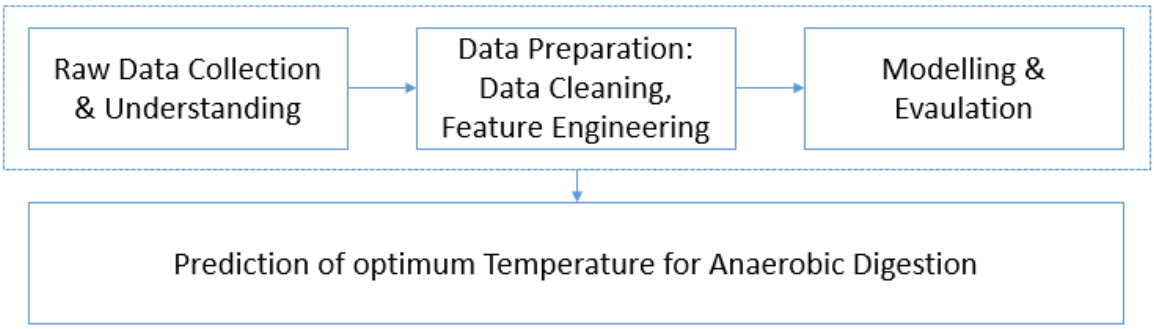

Figure 4. Overview of the methods.

During modelling, the features prepared in 'Data Preparation' are processed to finally make statements about patterns found in the data. Depending on the model, this methodology follows the train–test approach, wherein data are divided into training and test datasets to improve the model's success iteratively.

2.2.5. Evaluation

For the regression models, root mean square (RMSE) was used to identify the accuracy of the prediction.

Classification models can be divided into several categories such as binary, multiclass, multi-label and hierarchical classification. The confusion matrix is a two-dimensional (A, B) evaluation of numerical values. The confusion matrix is often used for cases with two classes, wherein they are appointed to true positives (TP), false positives (FP), true negatives (TN), and false negative (FN) [34].

In this study, the output was classified into three categories (A, B, C) that lead to multiclass classification, as in Figure 5.

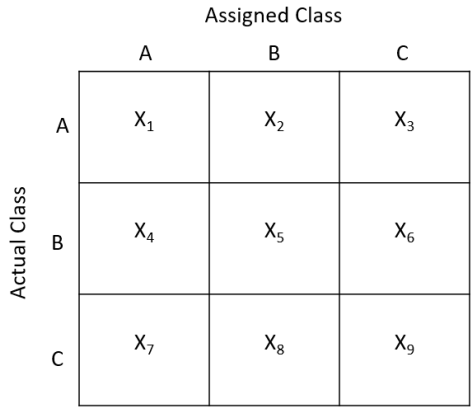

Figure 5. Confusion Matrix for multiclass classification.

In order to be able to evaluate an algorithm or its results, several metrics are needed, such as precision, recall, F1 score and support. All of these metrics will be used in the next section to compare the best results of the models.

Precision shows the correctness and accuracy of the model. As mentioned above, the confusion matrix has its true positives, true negatives, false positives and false negatives. Therefore, precision can be calculated with the following formula (Equation (4)) [34]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{X1}}{\text{X1} + \text{X2} + \text{X3}} \tag{4}$$

Recall, also referred to as sensitivity, is the ratio of correctly identified positive cases to all actual positive cases, i.e., the sum of "false negatives" and "true positives", as in Equation (5) [34].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{X1}}{\text{X1} + \text{X4} + \text{X7}} \tag{5}$$

Harmonic means of precision and recall can also be considered as F1-score, which can be considered a comparison of classification models as represented in Equation (6) [34].

$$\begin{aligned} \text{F1-score} = \quad & 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \\ & = \frac{2 \times 1}{2\text{X1} + \text{X2} + \text{X3} + \text{X4} + \text{X7}} \end{aligned} \tag{6}$$

The structural character of the training data can be indicated by using a support that implies the number of proceedings [34].

3. Results

As mentioned in Section 2.2.3, outliers were removed from the collected data, and since the difference between the scales of the variables could affect the models, the variables were standardized using the 'sklearn.preprocessing' library. Then, some of the missing data were replaced by the mean of the values in the same column, before starting the model building.

3.1. Comparison of the Prediction Models

The predictions of methane volume were performed using regression and classification models. For both regression and classification models, the dataset was split to a training and a test dataset. After splitting the dataset into a training and a test dataset, all of the models were trained using the training dataset. Finally, for the validation of the models, the test dataset was used for each algorithm.

As mentioned before, the target variable "class", which was converted to methane standard volume, was predicted by using machine learning algorithms.

After the model building and evaluation, among the regression models, the linear regression model, decision tree and random forest have RMES values of 246.96, 72.16, 93.91, respectively. The SVM algorithm gave the best accuracy of 0.93, followed by random forest (RF) with 0.89, kNN with 0.88 and decision tree with 0.86. Other models used for this dataset either gave much less precision or had overfitting. The accuracy of each model is shown in Table 2. All of the indicators for the accuracy of the model, such as "precision", "recall", "F1-score", and "support" were explained in Section 2.2.5—Evaluation using the confusion matrix.

An SVM model develops numerous strengths as a predictive method in machine learning that might also give the best accuracy out of the models trained. The reasons for the SVM model being the best model for the study could be the applicability of the model with linear and non-linear classification, and very complex non-linear partition surfaces that can be mapped by using additional dimensions and hyperplanes. However, an XG BOOT model gave a higher accuracy with 0.99, which can indeed be shown as an example of overfitting.

Table 2. Comparison of the classification models.

Metrics	Logistic Regression	Decision Tree	kNN	RF	SVM
Precision	0.62	0.86	0.88	0.89	0.93
Recall	0.64	0.87	0.87	0.89	0.92
F1-score	0.60	0.86	0.87	0.89	0.92
Support	189	36	142	0.12	142

Basically, overfitting means that the model is too specialized with the training dataset, so a very high model quality is achieved in the training dataset. Although there are several reasons and prevention methods for overfitting, it is probably about the size of the dataset in this study since the data were generated with several laboratory experiments.

A deeper investigation of the features and target value of methane production could be conducted by using feature importance. Since there was no ready-to-use feature importance template for our most accurate model, kNN, permutation feature importance was used with the features. The score for each feature is shown in Figure 6.

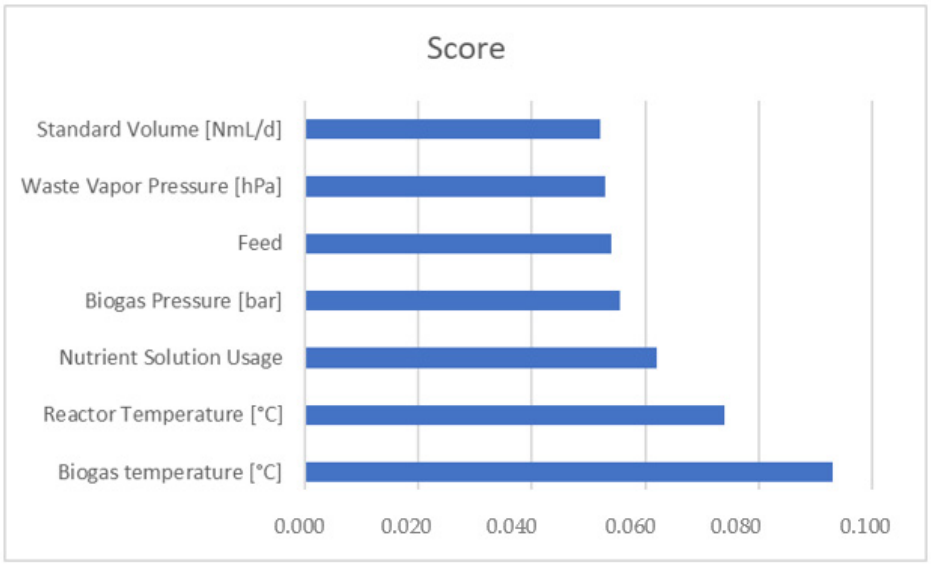

Figure 6. Permutation Feature Importance for kNN.

3.2. Results from Laboratory Tests

Results of the laboratory analyses were examined for each reactor specifically, since different OLRs and different temperature changing strategies were implemented in those reactors. The impact of temperature fluctuations was evaluated based on daily biomethane generation fluctuations, 7-day average biomethane generation, the methane content of biogas, pH, TS, VS, VOA/TIC, VFA and HCO_3^- content in the reactors. Due to the different OLRs, methane production in CSTR-1 and CSTR-2 was around 1000 NmL CH_4 d^{-1}, while it was about 2000 NmL CH_4 d^{-1} in reactors CSTR-3 and CSTR-4.

CSTR-1: The increase in temperature led to a decrease in biomethane generation and the methane content of biogas, as represented in Figure 7. Methane generation increased after keeping the temperature stable at 55 °C for three days. During the temperature decrease to 42 °C, when the temperature reached 50 °C, improvements in methane generation and methane content were achieved. Ten days' adaptation time was enough to reach the beginning conditions at 42 °C.

As represented by the laboratory analyses in Figure 8, the pH value in the reactor fluctuated between 7.30 and 7.90. The pH values were parallel with the temperature increases and decreases. The Highest VOA/TIC result (0.507) was obtained at 55 °C (after

seven days of operation at this temperature) and fluctuated between 0.150 and 0.507. A continuous decrease in TS resulted from the high TS content of the starting material of the reactor.

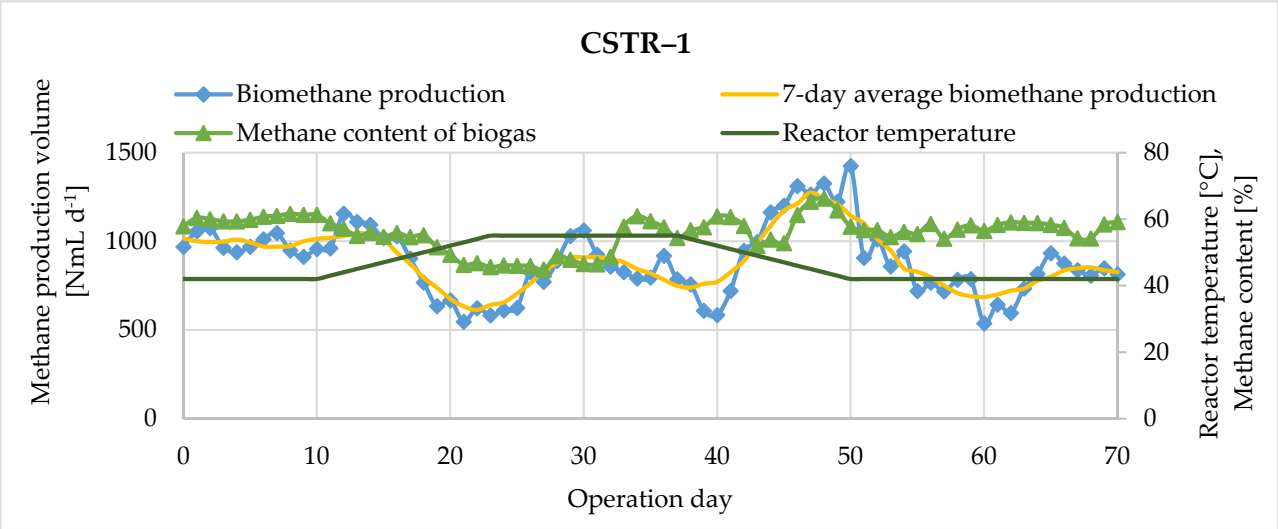

Figure 7. Operational results of CSTR-1.

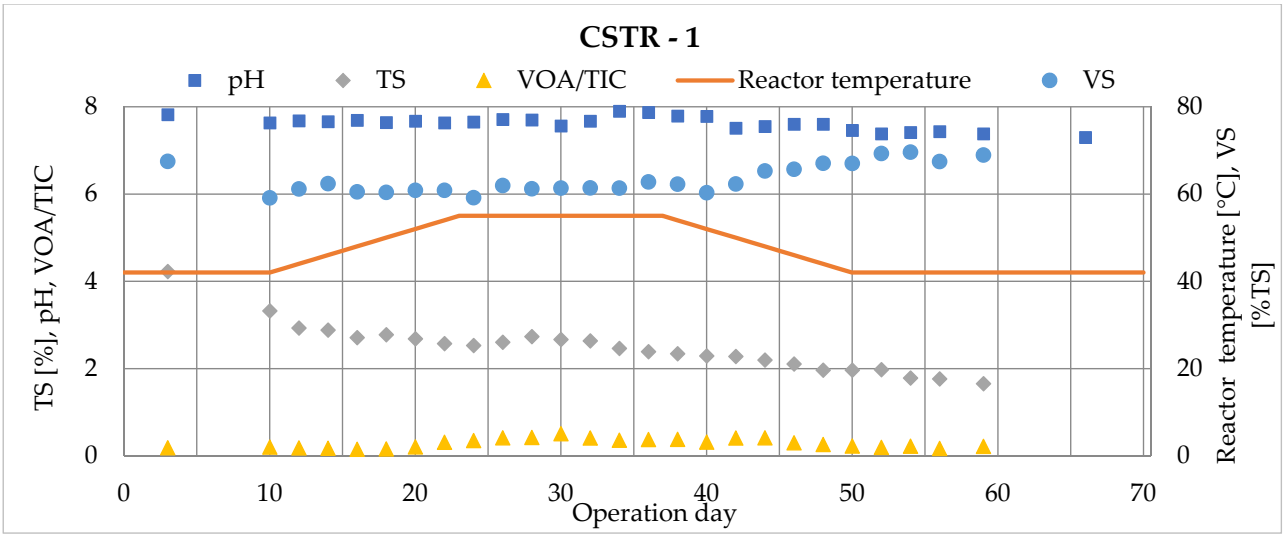

Figure 8. Results of laboratory analyses in CSTR-1.

During the temperature increase, biomethane generation was 9% lower than biomethane generation at 42 °C. On the other hand, the returning period to 42 °C resulted in a 5% higher biomethane generation than at the beginning, which can result from an improved hydrolysis rate at high temperatures.

CSTR-2: The difference between the second reactor and the first one was that the second reactor started with temperature decrease, as displayed in Figure 9. With the decrease in the temperature, the generation of methane fluctuated. Five days of operation at 29 °C resulted in the adaptation of the system to this temperature and increased methane generation. When the system reached 42 °C, five days were enough to observe acclimatization.

Figure 10 represents the results of the laboratory analysis for CSTR-2. The pH fluctuations were between 7.32 and 7.87, while VOA/TIC values ranged between 0.144 and 0.321. The impact of the temperature increase was more severe than the impact of the temperature

decreases on the system's stability. The required time for adaptation after reaching 42 °C was similar to the time required after the temperature increase.

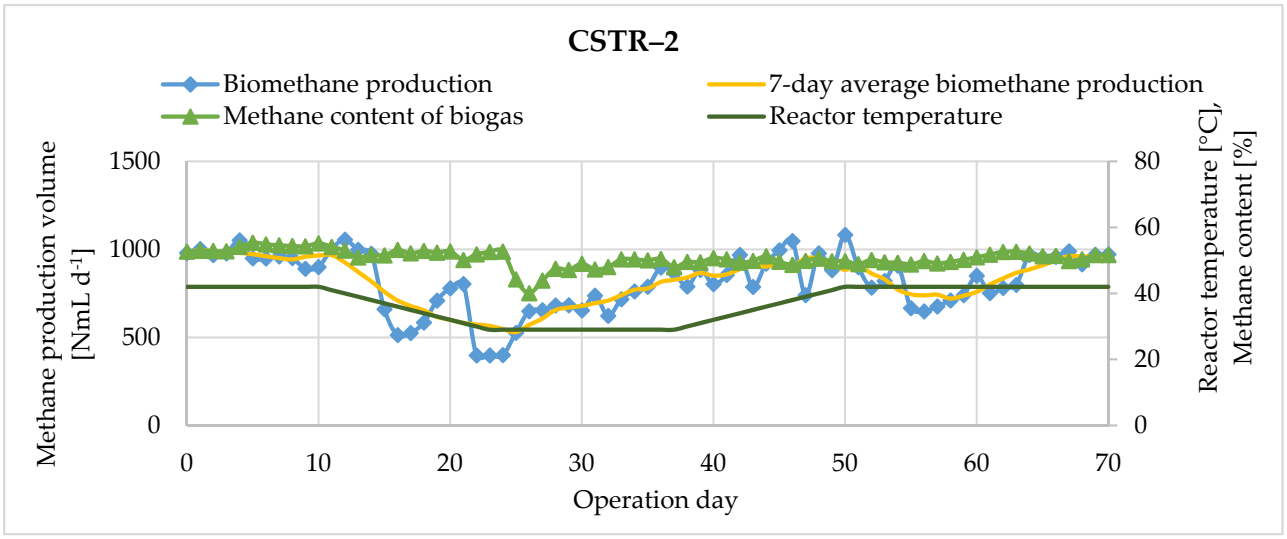

Figure 9. Operational results of CSTR-2.

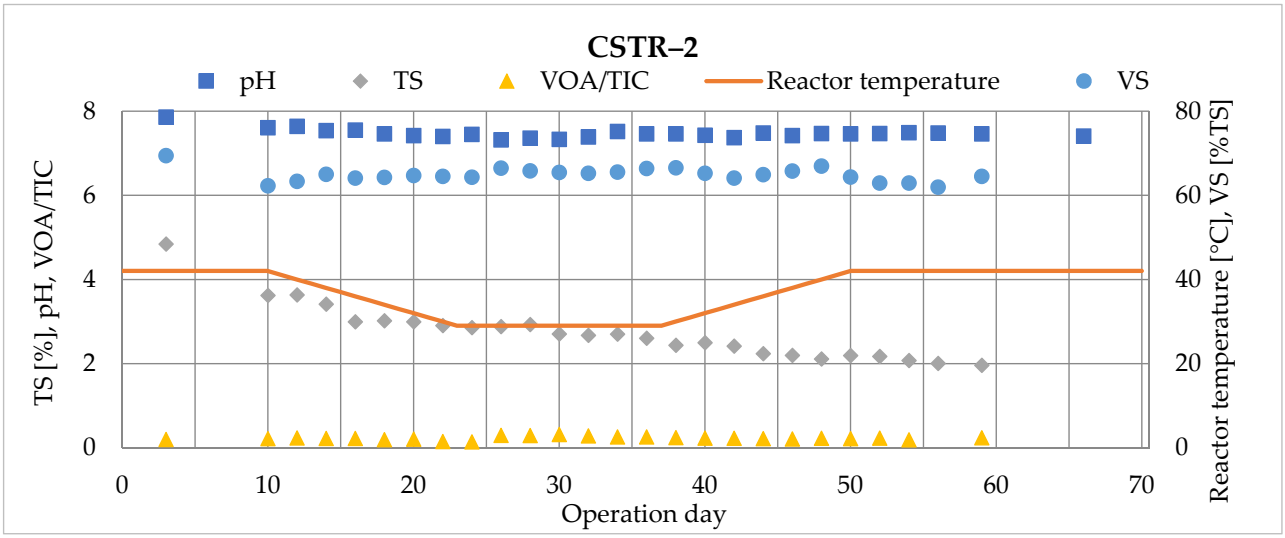

Figure 10. Results of laboratory analyses in CSTR-2.

During the temperature decrease, 24% less biomethane was obtained from the beginning conditions. The increase in temperature back to the beginning temperature improved the biomethane generation.

CSTR-3: While the OLR of CSTR-3 was two times higher than CSTR-1 and CSTR-2, the same temperature increase scenarios were implemented as in CSTR-1, as represented in Figure 11. The temperature increase led to increased biomethane generation, which decreased after ten days. Operation at 55 °C resulted in the first increase and decrease in the biomethane generation. Although high fluctuations in the laboratory results were observed in this reactor (see Figure 12), an adaptation of the system to the beginning conditions (after returning to 42 °C) required a shorter time than the time needed in CSTR-1. Overall, when the temperature increased from 42 °C to 55 °C, the obtained average biomethane amount was 10% less than in the beginning conditions. During operation at 55 °C, this decrease was 33%.

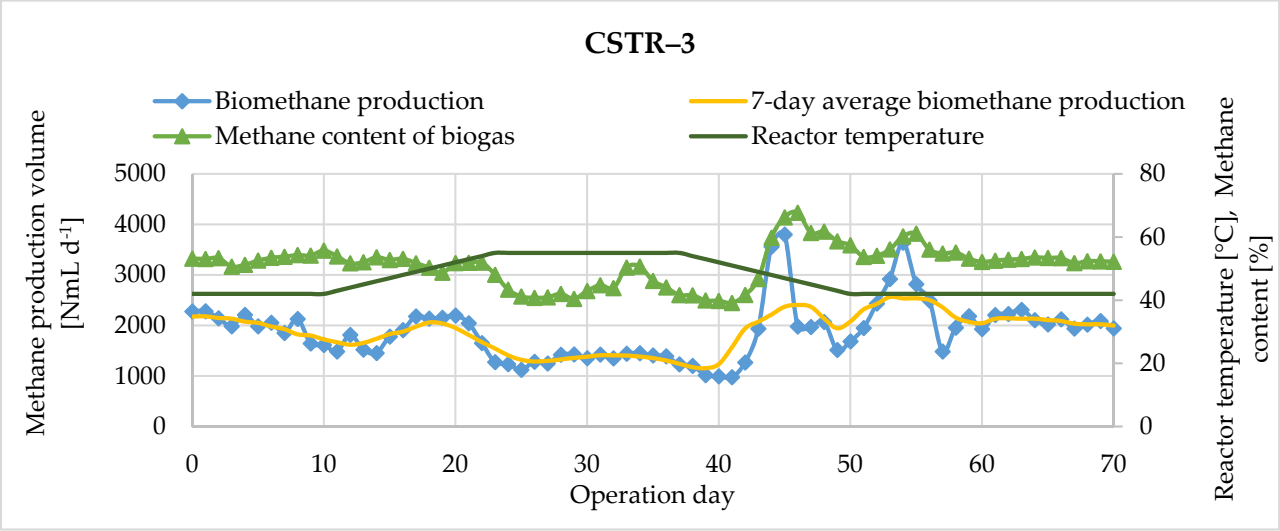

Figure 11. Operational results of CSTR-3.

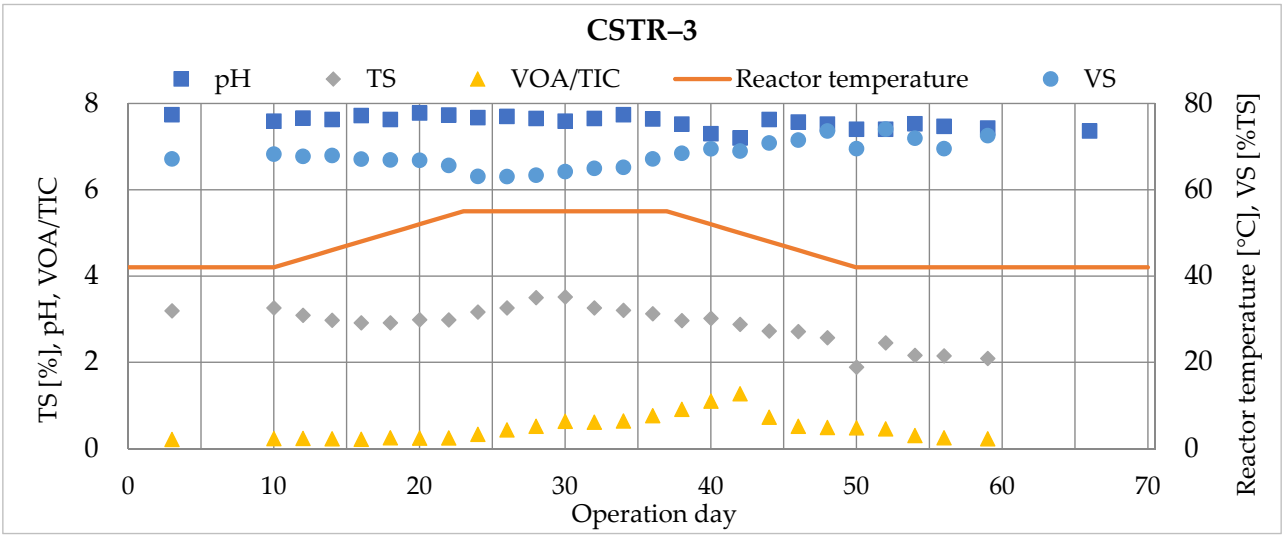

Figure 12. Laboratory analyses results of CSTR-3.

The pH value fluctuated between 7.20 and 7.78 within this period (see Figure 12). The VOA/TIC results were relatively higher than CSTR-1 due to the higher organic load in this reactor. Although extremely high VOA/TIC results were obtained from this reactor, the recovery time needed at 42 °C was shorter than the time needed in CSTR-1. For CSTR-1, it is likely that an increase in the temperature increased the hydrolysis rate of the substrate.

CSTR-4: The same temperature change strategy implemented in CSTR-2 was implemented in CSTR-4 with higher OLR. A decrease in temperature resulted in decreased biomethane production, but almost stable methane content was obtained during the temperature changes, as represented in Figure 13. Biomethane generation started increasing after ten days of operation at 29 °C. During temperature changes, fluctuations in biomethane generation were lower than the fluctuations in the other reactors.

The pH value fluctuations ranged between 7.31 and 7.72, as represented in Figure 14, showing a parallel relation to the temperature changes. VOA/TIC fluctuated between 0.087 and 0.273, with the highest value on day 28 when the temperature was kept stable at 29 °C for five days. All in all, a 2% decrease in biomethane generation was obtained during the decrease in temperature. Increasing the temperature back to 42 °C caused higher biomethane generation (approximately 6% higher than the generation at the beginning).

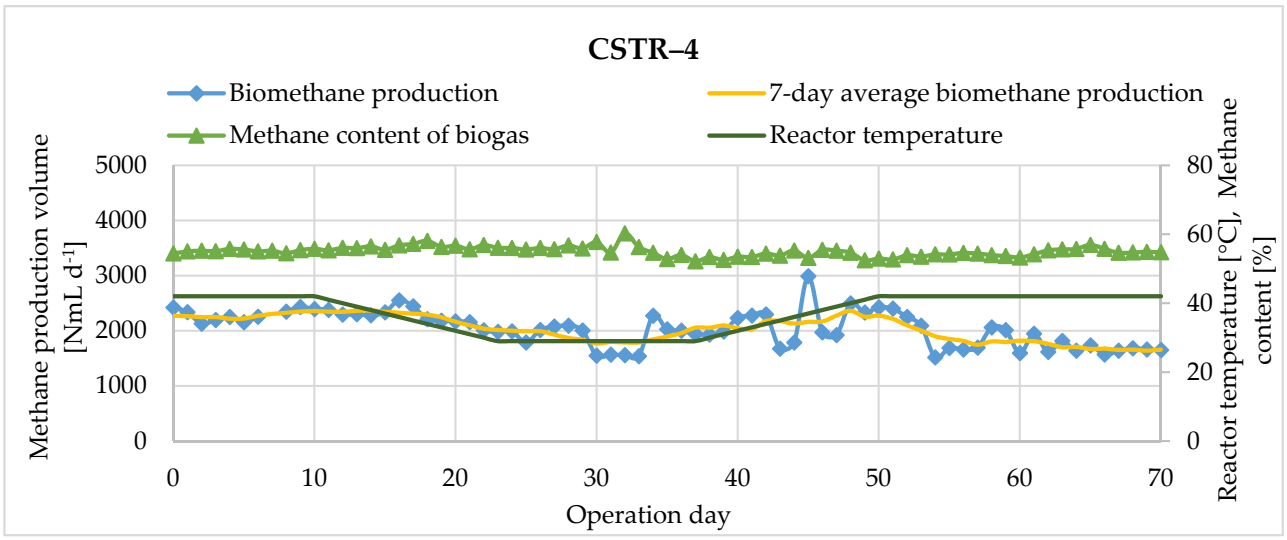

Figure 13. Operational results of CSTR-4.

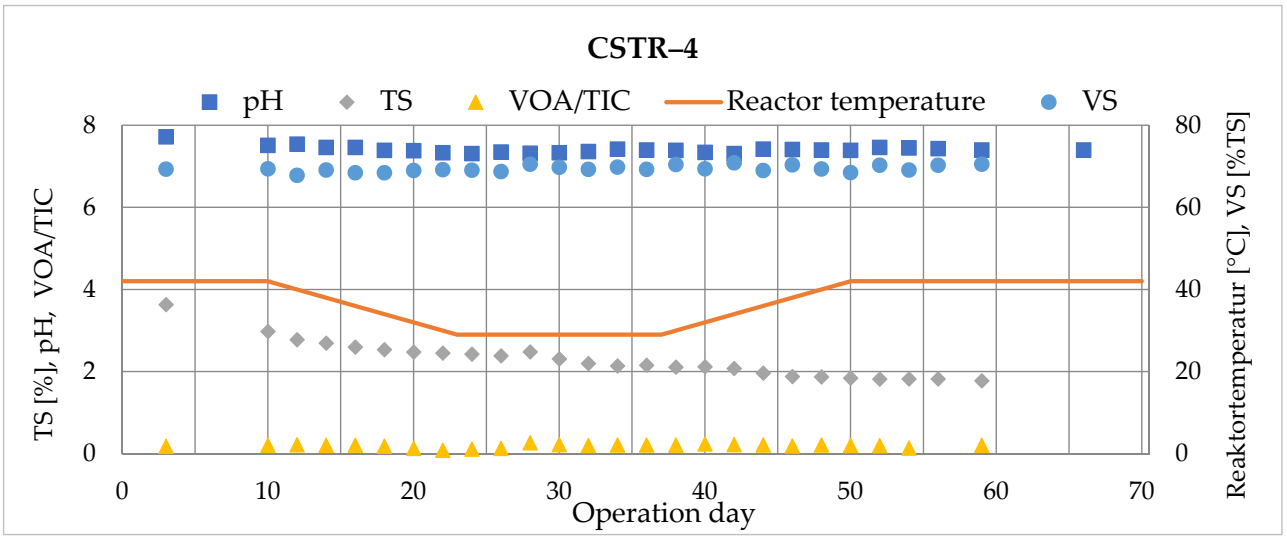

Figure 14. Laboratory analyses results of CSTR-4.

The results of the VFA and HCO_3^- analyses are represented in Figure 15. High fluctuations in VFA content were obtained from reactors CSTR1 and CSTR3. The accumulation of VFAs showed that there was a disruption in the stages of the process, where the balance of material recovery could not be supplied. On the other hand, HCO_3^- concentration did not fluctuate severely in the four reactors.

Overall, even after high fluctuations in the process parameters, recovery of the process was possible after some time. Due to temperature fluctuations happening in four reactors, increases in the VFAs and HCO_3^- were observed, as represented in Figure 15. In particular, temperature increases in CSTR-1 and CSTR-3 led to the destruction of the process, and eventually the accumulation of VFAs. The results of laboratory analyses showed that it is possible to change the temperature regime in the studied anaerobic digestion process without having an irreversible impact on the process stability in the used reactors. The minimum effect on the process was in CSTR-4 regarding the process parameters, while the shortest recovery time was required for CSTR-3. In general, after a specific time of acclimation, each process recovered from the changes.

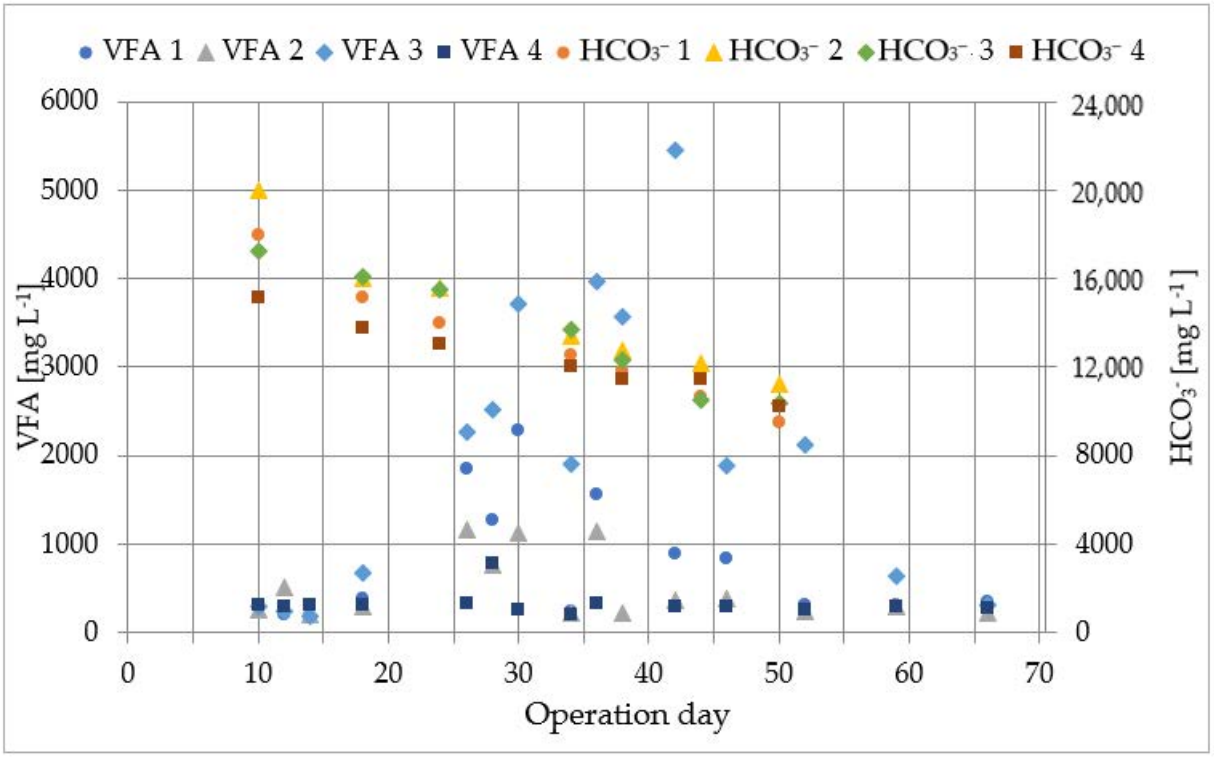

Figure 15. Results of VFA and HCO_3^- analyses in all reactors.

4. Discussion

Our study addressed the question of whether predictive analytics methods such as machine learning can be used for biogas operation, by predicting the optimal operating temperature of anaerobic digestion. Due to the wide variety and high degree of complexity of biological systems, different algorithms can be found in the literature for different studies, with varying degrees of accuracy. Therefore, it would be best to develop and create a separate model for each system and environment. This work also aimed to provide a blueprint for how to create a machine learning model, from data gathering to evaluation, that can be modified for other systems if required [24,27,35]. In order to build a machine learning model, several experiments were conducted with different operational temperatures at a laboratory scale, and the collected data were used in different regression and classification models. Conducted models showed some meaningful accuracy for predicting methane production. From the regression models, the decision tree had the best accuracy with an RMSE value of 72.16, and as a classification method, the model with the support vector machine had the best accuracy with 0.93 Precision, calculated by a confusion matrix as explained in Section 2.2.5 Evaluation.

The possible reason for the lower accuracies of the other models could be the degree of nonlinearity of the models; this could also possibly be the reason for the difference between the numerical prediction of methane production with the regression analysis, and the categorical prediction with the classification models. Since the models were built with data gathered in environmental conditions, it was easier to keep the data and the parameters stable. In a real-world application with an industrial-scale biogas plant, the operational parameters could even show daily base fluctuations. This dynamic behavior of the system could necessitate more frequent maintenance of the model. After having the best accuracy with the SVM model, the best tuning-parameters were searched with the help of the meta-estimator Gridsearch function in python.

As mentioned above, statistical methods have already been used in the biological science for process optimization. Among these statistical methods, some modern approaches

such as predictive analytics and machine learning have been suggested for better operation, with the type of biowaste as the substrate and the total carbon amount as an input. In our study, the feasibility of the usage of machine learning was investigated with the operational parameter of temperature during anaerobic digestion.

Our findings offer a novel perspective on the feasibility of predictive statistical on biogas technologies for better operation performance. Applying these models on an industrial scale will certainly also require the introduction of some additional analyses and methods at each stage of the modelling. Furthermore, considering the possible variety of organizational parameters in anaerobic digestion, it is complicated to have an overview of the whole process, especially for predictive analytics. Therefore, it is of tremendous importance to ensure consistency during modelling and at every stage of the abovementioned analysis method, such as data acquisition, data interpretation, data preparation, model building, and evaluation.

This appropriate approach to analysis could be achieved with the help of a common understanding of the project goal for all stakeholders, defining the desired output for the project; this includes: in-scope and out-of-scope issues; ensuring the robustness of the measurement methods, starting with the correct definition of the output metrics; operational definition; data types; and analysis of the measurement systems to be able to analyze the process capability before starting the analysis. Furthermore, for the analysis of the collected data, it is possible to support the whole modelling process with some laboratory analyses on strategically selected samples.

Our study has two main limitations. The first limitation was surely the challenge of conducting statistic analyses with the data gathered from a living organism. However, as mentioned earlier, the severity of this can be overcome with the help of sufficient monitoring of the process and good quality data. Furthermore, in this study, sufficient time was given after changing the temperature in each case. Furthermore, VFA and $HCO_3{}^{-1}$ analyses were carried out to ensure process stability. Another limitation is the relatively small sample size, as the experiments were conducted in a laboratory setting and in limited time. Nevertheless, the results were representative despite the amount of data.

We investigated whether machine learning could be used in anaerobic digestion, and our findings confirm that modern statistical approaches can be useful under some conditions in the biological system. Unquestionably, more research should be conducted in this area with various of types and combinations of the operational parameter for more robust modelling. With the sufficient usage of machine learning, it might be possible to create a certain level of automation and self-deciding systems for process optimization in biological systems such as anaerobic digestion.

5. Conclusions

In this study, different machine learning models were trained to predict biogas output under different intensities of feeding and different temperature variations. This model could be used to predict the effect of temperature variations on process efficiency, enabling the real-time monitoring and control strategy for biogas technology. As mentioned in the discussion, it is possible to extend the application of predictive analytics for process optimization at all possible scales for anaerobic digestion. In order to be able to gain better accuracy from the models, each phase of the methodology, from the beginning problem-definition to the end evaluation, must be organized, structural and goal-oriented.

Author Contributions: Conceptualization, S.Ö.C. and S.C.; methodology, S.C.; software, S.C.; validation, S.Ö.C.; formal analysis, S.Ö.C.; writing—original draft preparation, S.C. and S.Ö.C.; writing—review and editing, S.Ö.C., S.C. and K.K.; visualization, S.C.; supervision, S.Ö.C. and K.K.; project administration, K.K.; funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Publishing fees were supported by Funding Programme "Open Access Publishing"of the Hamburg University of Technology. We would like to thank the German Academic Exchange Service (DAAD) for their scholarship to Senem Önen Cinar.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Energy Agency. Germany 2020—Energy Policy Review. Available online: https://www.iea.org/reports/germany-2020 (accessed on 26 January 2022).
2. FNR. Flexibilisierung von Biogasanlagen 2018. Available online: https://mediathek.fnr.de/flexbroschuere.html (accessed on 26 January 2022).
3. Cinar, S.; Cinar, S.O.; Wieczorek, N.; Sohoo, I.; Kuchta, K. Integration of artificial intelligence into biogas plant operation. *Processes* **2021**, *9*, 85. [CrossRef]
4. Drosg, B. *Process Monitoring in Biogas Plants*; IEA Bioenergy: Paris, France, 2013; ISBN 1910154024.
5. Nsair, A.; Onen Cinar, S.; Alassali, A.; Abu Qdais, H.; Kuchta, K. Operational Parameters of Biogas Plants: A Review and Evaluation Study. *Energies* **2020**, *13*, 3761. [CrossRef]
6. Cruz, I.A.; de Melo, L.; Leite, A.N.; Melquiades Sátiro, J.V.; Santos Andrade, L.R.; Torres, N.H.; Cabrera Padilla, R.Y.; Bharagava, R.N.; Tavares, R.F.; Romanholo Ferreira, L.F. A new approach using an open-source low cost system for monitoring and controlling biogas production from dairy wastewater. *J. Clean. Prod.* **2019**, *241*, 118284. [CrossRef]
7. Deng, L.; Liu, Y.; Wang, W. *Biogas Technology*; Springer Singapore: Singapore, 2020; ISBN 978-981-15-4939-7.
8. Tian, G.; Yang, B.; Dong, M.; Zhu, R.; Yin, F.; Zhao, X.; Wang, Y.; Xiao, W.; Wang, Q.; Zhang, W. The effect of temperature on the microbial communities of peak biogas production in batch biogas reactors. *Renew. Energy* **2018**, *123*, 15–25. [CrossRef]
9. Westerholm, M.; Müller, B.; Isaksson, S.; Schnürer, A. Trace element and temperature effects on microbial communities and links to biogas digester performance at high ammonia levels. *Biotechnol. Biofuels* **2015**, *8*, 1–19. [CrossRef]
10. Chae, K.J.; Jang, A.; Yim, S.K.; Kim, I.S. The effects of digestion temperature and temperature shock on the biogas yields from the mesophilic anaerobic digestion of swine manure. *Bioresour. Technol.* **2008**, *99*, 1–6. [CrossRef] [PubMed]
11. Pap, B.; Györkei, Á.; Boboescu, I.Z.; Nagy, I.K.; Bíró, T.; Kondorosi, É.; Maróti, G. Temperature-dependent transformation of biogas-producing microbial communities points to the increased importance of hydrogenotrophic methanogenesis under thermophilic operation. *Bioresour. Technol.* **2015**, *177*, 375–380. [CrossRef]
12. Membere, E.; Sallis, P. Effect of temperature on kinetics of biogas production from macroalgae. *Bioresour. Technol.* **2018**, *263*, 410–417. [CrossRef]
13. Bavutti, M.; Guidetti, L.; Allesina, G.; Libbra, A.; Muscio, A.; Pedrazzi, S. Thermal stabilization of digesters of biogas plants by means of optimization of the surface radiative properties of the gasometer domes. *Energy Procedia* **2014**, *45*, 1344–1353. [CrossRef]
14. Hubert, C.; Steiniger, B.; Schaum, C.; Michel, M.; Spallek, M. Variation of the digester temperature in the annual cycle–using the digester as heat storage. *Water Pract. Technol.* **2019**, *14*, 471–481. [CrossRef]
15. Terradas-Ill, G.; Pham, C.H.; Triolo, J.M.; Martí-Herrero, J.; Sommer, S.G. Thermic model to predict biogas production in unheated fixed-dome digesters buried in the ground. *Environ. Sci. Technol.* **2014**, *48*, 3253–3262. [CrossRef] [PubMed]
16. Powell, D.; Lundeby, S.; Chabada, L.; Dreyer, H. Lean Six Sigma and environmental sustainability: The case of a Norwegian dairy producer. *Int. J. Lean Six Sigma* **2017**, *8*, 53–64. [CrossRef]
17. Boe, K. Online Monitoring and Control of the Biogas Process. 2006. Available online: https://www.osti.gov/etdeweb/biblio/20833720 (accessed on 26 January 2022).
18. Kara, S.; Mueller, J.J.; Liese, A. Online analysis methods for monitoring of bioprocesses. *Chem. Today* **2011**, *29*, 2.
19. Lee, C.K.; Cao, Y.; Ng, K.H. Big data analytics for predictive maintenance strategies. In *Supply Chain Management in the Big Data Era*; IGI Global: Hershey, PA, USA, 2017; pp. 50–74.
20. Mauky, E.; Weinrich, S.; Jacobi, H.-F.; Nägele, H.-J.; Liebetrau, J.; Nelles, M. Demand-driven biogas production by flexible feeding in full-scale–Process stability and flexibility potentials. *Anaerobe* **2017**, *46*, 86–95. [CrossRef] [PubMed]
21. Wahmkow, C.; Knape, M.; Konnerth, E. Biogas Intelligence—Operate biogas plants using Neural Network and Fuzzy logic. In Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton, AB, Canada, 24–28 June 2013; pp. 1483–1488, ISBN 978-1-4799-0348-1.
22. Manu, D.S.; Thalla, A.K. Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. *Appl Water Sci* **2017**, *7*, 3783–3791. [CrossRef]
23. Vanti, C.V.M.; Leite, L.C.; Batista, E.A. Monitoring and control of the processes involved in the capture and filtering of biogas using FPGA embedded fuzzy logic. *IEEE Lat. Am. Trans.* **2015**, *13*, 2232–2238. [CrossRef]
24. Wang, L.; Long, F.; Liao, W.; Liu, H. Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. *Bioresour. Technol.* **2020**, *298*, 122495. [CrossRef]
25. Nourani, V.; Elkiran, G.; Abba, S.I. Wastewater treatment plant performance analysis using artificial intelligence–An ensemble approach. *Water Sci. Technol.* **2018**, *78*, 2064–2076. [CrossRef]

26. Bolette, D.H.; Jamshid, T.; Christian, A.T.; Rasmus, J.; Thomas, B.M.; David, J.G. *Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model*; International Conference on Computational Science and Its Applications; Springer: Berlin/Heidelberg, Germany, 2020.
27. De Clercq, D.; Jalota, D.; Shang, R.; Ni, K.; Zhang, Z.; Khan, A.; Wen, Z.; Caicedo, L.; Yuan, K. Machine learning powered software for accurate prediction of biogas production: A case study on industrial-scale Chinese production data. *J. Clean. Prod.* **2019**, *218*, 390–399. [CrossRef]
28. Kaltschmitt, M.; Hartmann, H.; Hofbauer, H. *Energie aus Biomasse*; Springer: Berlin/Heidelberg, Germany, 2009; ISBN 978-3-540-85094-6.
29. *The Biogas Handbook*; Elsevier: Amsterdam, The Netherlands, 2013; ISBN 9780857094988.
30. Zhang, X.-D. Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 223–440.
31. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-7137-0.
32. Bonaccorso, G. *Machine Learning Algorithms*; Packt Publishing Ltd.: Birmingham, UK, 2017; ISBN 1785884514.
33. Ekström, M.; Esseen, P.-A.; Westerlund, B.; Grafström, A.; Jonsson, B.G.; Ståhl, G. Logistic regression for clustered data from environmental monitoring programs. *Ecol. Inform.* **2018**, *43*, 165–173. [CrossRef]
34. Xu, J.; Zhang, Y.; Miao, D. Three-way confusion matrix for classification: A measure driven view. *Inf. Sci.* **2020**, *507*, 772–794. [CrossRef]
35. Cruz, I.A.; Chuenchart, W.; Long, F.; Surendra, K.C.; Andrade, L.R.S.; Bilal, M.; Liu, H.; Figueiredo, R.T.; Khanal, S.K.; Ferreira, L.F.R. Application of machine learning in anaerobic digestion: Perspectives and challenges. *Bioresour. Technol.* **2022**, *345*, 126433. [CrossRef] [PubMed]