Hindawi
International Journal of Chemical Engineering
Volume 2023, Article ID 2871769, 11 pages
https://doi.org/10.1155/2023/2871769

Research Article

Estimation of Acetic Acid Concentration from Biogas Samples Using Machine Learning

Lingga Aksara Putra, Bernhard Huber, and Matthias Gaderer

Technical University of Munich, Professorship of Regenerative Energy Systems, Schulgasse 16 94315, Germany

Correspondence should be addressed to Lingga Aksara Putra; lingga_aksara.putra@tum.de

Received 2 December 2022; Revised 10 January 2023; Accepted 25 January 2023; Published 22 February 2023

Academic Editor: Achim Kienle

In a biogas plant, the acetic acid concentration is a major component of the substrate as it determines the pH value, and this pH value correlates with the volume of biogas produced. Since it requires specialized laboratory equipment, the concentration of acetic acid in a biogas substrate cannot be measured on-line. The project aims to use NIR sensors and machine learning algorithms to estimate the acetic acid concentration in a biogas substrate based on the measured intensities of the substrate. As a result of this project, it was possible to determine whether the acetic acid concentration in a biogas substrate is higher or lower than 2 g/l using machine learning models.

1. Introduction

Germany introduced the first EEG (Renewable Energy Sources Act) in 2000 [1], which encouraged the expansion of wind, solar, and biogas power plants. A 20 year feed in tariff guaranteed by the government made these technologies financially attractive for many operators. In the coming years, the first plants will drop out of the EEG and will have to compete in the free electricity market, unless new political regulations are put in place for them.

The first step toward demand-oriented power generation from biogas plants was made possible with the introduction of the flexibility benefit for additional CHP capacity under the EEG 2012. In an electricity market with a high proportion of volatile forms of generation such as wind power and photovoltaics, this can be an important contribution to balancing power and residual load coverage. To further increase flexibility without having to invest in additional gas storage capacity, it would be conceivable to change biogas production from constant to demand-oriented by employing new feeding concepts [2–4].

Obtaining demand-oriented biogas production is challenging due to the lack of measured states in the biogas digester. An important value of the substrate in the digester is the acetic acid concentration, which determines the pH of the substrate, and a change in pH can negatively impact biogas production [5]. It is impractical to measure the acetic acid concentration of a biogas substrate directly in a biogas plant as this must be done in the laboratory usually using either high-performance liquid chromatography (HPLC) or gas chromatography (GC) [6].

In light of the above points, this paper aims to investigate whether machine learning can be used to measure the acetic acid concentration directly and thus save time and resources since no further laboratory work is needed. On-line measurement of acetic acid concentration in a biogas digester represents a significant step forward to achieving demand-oriented biogas production.

2. Review of Literature

The two main approaches to achieving demand-oriented biogas production are to increase gas storage capacity or to implement flexible feeding strategies. Physical, legal, and economic constraints restrict the ability to expand gas storage capacities. This suggests that the second method may be more preferred because it allows biogas production to be adjusted to gas and electricity demand by varying substrates

and feeding amounts [7]. Additionally, a biogas plant has been evaluated for its ability to handle different feeding management strategies, such as feeding less frequently without compromising its long-term stability [8].

Feeding management strategies can be abstracted as a feedback control problem. To solve a control problem, a system model for dynamic simulation is required. Modeling a biogas plant can be done using a first-principle mathematical model [9–11] or by using machine learning techniques based on analyzed data [12–14]. Among both methods, a first-principal-based mathematical model enables controlling the plant with an advanced feedback controller such as a model predictive controller [15, 16] or a PI controller [17].

Due to the complex chemical processes involved in a biogas plant, creating a first-principle-based mathematical model can be extremely challenging. Additionally, a first-principle model requires a variety of laboratory measurements to ensure that the model parameters are reliable, which is time-consuming and resource-intensive. Therefore, the control accuracy of a first-principle-based model of a plant is strongly influenced by the modeling complexity of the physical-chemical process and the accuracy of the laboratory measurements.

As opposed to the first-principle-based mathematical model, the data-based machine learning models are derived only by analyzing a large number of measurements. Machine learning algorithms can identify a direct relationship between data groups to build a black-box model for accurate prediction. Considering the above discussion, this paper focuses on the data-based machine learning model.

At the beginning of the rise of the machine learning era, the main focus of machine learning or artificial intelligence was on the development of computer vision algorithms [18], problem solving for computational applications [19], natural language processing [20, 21], and virtual assistants [22, 23]. Currently, machine learning can also be applied in chemistry [24] as well as renewable energy [25].

A machine learning model of a biogas plant does not require a complex mathematical equation, but rather a great deal of training data, which come from sensors installed in the plant, such as temperature, pressure, gas flow rate, and near-infrared (NIR) sensors.

NIR spectroscopy is capable to determine the total nitrogen (TN), organic carbon (OC), and moisture content (MC) of soil [26, 27]. It is also possible to use NIR sensors more widely, for example, in the early detection of insects [28] as well as in the food and beverage industry [29, 30].

In light of these findings, it is reasonable to assume that NIR sensors can be used to estimate acetic acid concentration in biogas substrates [31–33]. Similarly, machine learning was used in the aforementioned studies to estimate the acetic acid concentration. However, the target concentration of 12 g/l [31] was too high because the acetic acid concentration in the digester must be less than 2 g/l [34]. It is therefore necessary to predict if the concentration is below or above 2 g/l.

3. Research Methodology

3.1. Laboratory Experiment. Throughout this project, samples were continuously taken from the biogas plant Grub [35] to be analyzed in the laboratory. Sample preparation has particular importance in laboratory analytics, and a major challenge was determining the proper sample preparation procedure to guarantee reproducible measurements and minimize sample preparation interference. The following steps (Figure 1) were used to prepare the sample after extensive research was conducted.

Since acetic acid influences not just the sample's acetic acid content but also its pH level, sodium acetate was added to the sample in place of acetic acid. The primary motivation is that pH fluctuations have an impact on the measured intensity of NIR sensors [36, 37]. The addition of sodium acetate to the sample only changes the acetic acid concentration and has no effect on the pH value. As a result, the addition of sodium acetate ensures that the machine learning model is calibrated solely on acetic acid concentrations.

Spectral Engines provided 4 NIR sensors for this project. Due to their small size and ease of installation, these sensors are used in this study instead of traditional lab NIR spectrometers. Table 1 presents their types and wavelength ranges. Overall, the sensors can measure wavelengths between 1100 and 2150 nm with a wavelength resolution (step size) of 2 nm.

The cuvette was initially measured horizontally on the NIR sensors. However, subsequent measurements confirmed that the intensity measured was not reproducible due to air bubbles in the cuvette. The cuvette may not always be filled with the sample, and there may be air bubbles. As the samples are horizontally placed, air bubbles will be located in the center of the cuvette, disrupting the intensity measurements. To solve the problem, a 3D-printed fixture was created to measure the cuvette vertically (Figure 2).

Glass cuvettes are required when measuring intensities in the NIR spectrum, due to the incompatibility of plastic cuvettes with this measurement.

3.2. Measurement. In this study, the NIR sensors were calibrated to classify samples according to their acetic acid concentration. Initially, the acetic acid concentration is present in samples from the biogas plant, and the measured concentration arises from adding sample concentration and sodium acetate concentration.

The first step was to determine whether the sodium acetate solution changed the measured acetic acid concentration of the samples. Four samples were prepared and measured using GC. The first sample (sample 0, Figure 3) was the zero sample without the addition of sodium acetate. The remaining three samples (samples 1–3, Figure 3) were fortified with sodium acetate so that the addition was about 2 g/l acetic acid. Gas chromatography was then used to measure acetic acid concentration in the samples.

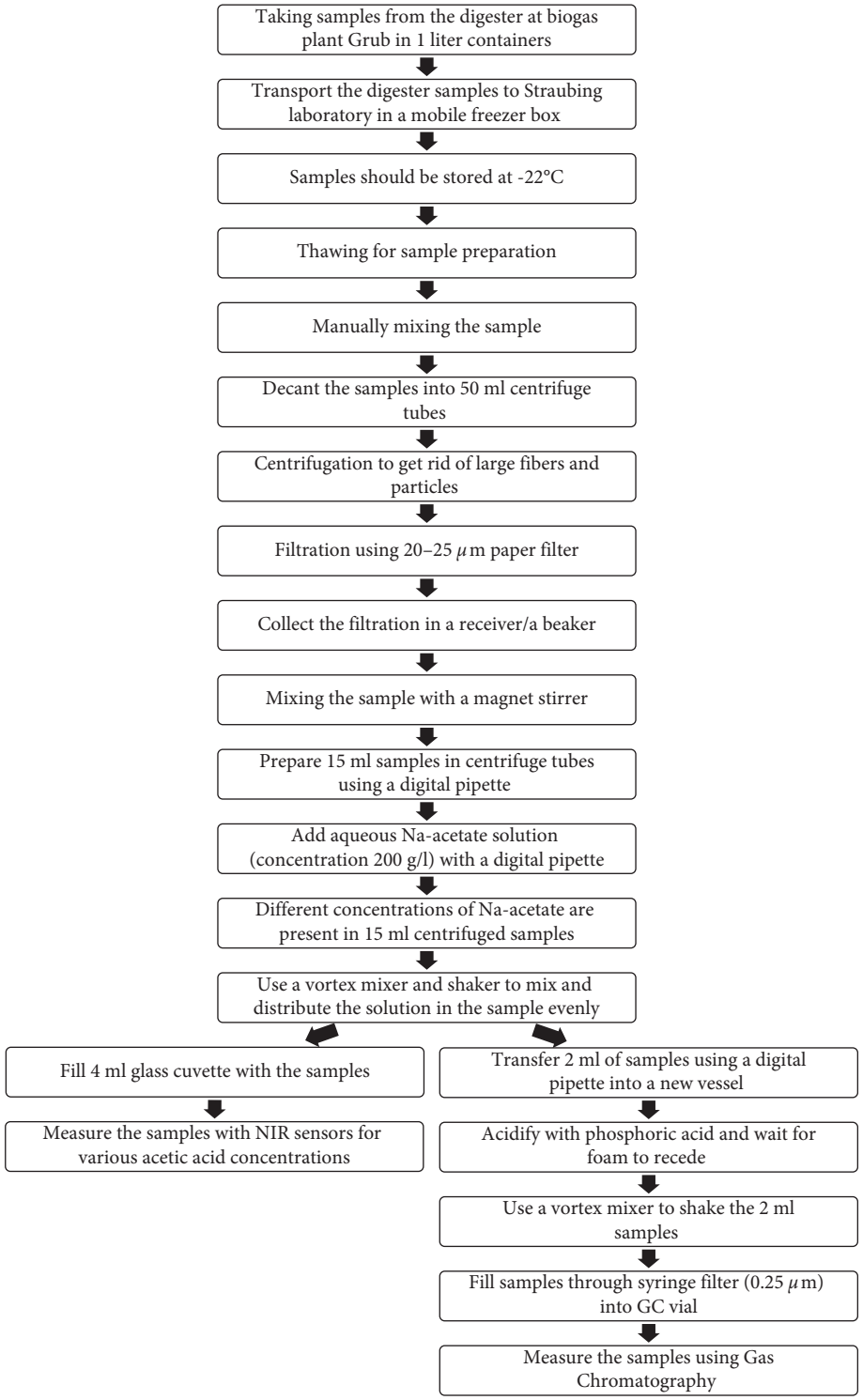


Figure 1: Detailed steps for preparing the samples.

The magenta line with yellow markers represents the theoretically calculated acetic acid concentration of the samples, and the green line with blue markers represents the measured acetic acid concentration of the samples from the gas chromatography. Analyzing the two lines, it is clear that they are highly correlated. This result guided the preparation of 264 samples for the first calibration of the NIR sensors. These samples consisted of the following:

Table 1: Sensor types and wavelength ranges.

NIR sensor	Wavelength range (nm)
S1.4	1100–1350
S1.7	1350–1650
S2.0	1550–1950
S2.2	1750–2150

(a) 133 samples with acetic acid concentrations <2 g/l (class 0)

(b) 131 samples with acetic acid concentrations >2 g/l (class 1)

As shown in Figure 4, the measured values for these 264 samples were obtained using 4 NIR sensors (S1.4, S1.7, S2.0, and S2.2).

Green lines represent measurements of 133 samples with acetic acid concentrations less than 2 g/l, and magenta lines represent the measurements of the 131 samples with acetic acid concentrations greater than 2 g/l. The use of machine learning is beneficial in this case because it is impossible to determine which features are important for classification by hand.

4. Data Analysis

4.1. Maximal Distance-to-Noise Ratio (MNR). It is important to first introduce the MNR, an important ratio found only in this study. The MNR, in principle, is equivalent to the signal-to-noise ratio (SNR). As opposed to comparing the noise and the signal, the MNR is calculated by comparing the noise with the maximal distance of intensity from measurements of 264 samples. The MNR can be calculated as follows.

Let $x^i_{w,t}$ be the measurement for sample i at wavelength w at time t. First, the mean value $\overline{x}$ and the variance σ are calculated for each sample at each wavelength during the measurement periods $\forall t \in \{1, \ldots, N\}$.

$$\overline{x}^i_w = \frac{\sum_{t=1}^{N} x^i_{w,t}}{N} = \begin{bmatrix} \overline{x}^1_1 & \cdots & \overline{x}^1_W \\ \vdots & \ddots & \vdots \\ \overline{x}^{264}_1 & \cdots & \overline{x}^{264}_W \end{bmatrix}, \quad (1)$$

$$\sigma^i_w = \sqrt{\frac{\sum_{t=1}^{N} \left| x^i_{w,t} - \overline{x}^i_w \right|^2}{N-1}} = \begin{bmatrix} \sigma^1_1 & \cdots & \sigma^1_W \\ \vdots & \ddots & \vdots \\ \sigma^{264}_1 & \cdots & \sigma^{264}_W \end{bmatrix}. \quad (2)$$

Find the maximum standard deviation across all samples $\forall i \in \{1, \ldots, 264\}$ using (2):

$$\max \sigma_w = \left[\max_{i=1\ldots264} \sigma^i_1 \quad \max_{i=1\ldots264} \sigma^i_2 \quad \cdots \quad \max_{i=1\ldots264} \sigma^i_W \right]. \quad (3)$$

The same method is used for calculating the maximum and minimum average values for (1):

$$\max \overline{x}_w = \left[\max_{i=1\ldots264} \overline{x}^i_1 \quad \max_{i=1\ldots264} \overline{x}^i_2 \quad \cdots \quad \max_{i=1\ldots264} \overline{x}^i_W \right], \quad (4)$$

$$\min \overline{x}_w = \left[\min_{i=1\ldots264} \overline{x}^i_1 \quad \min_{i=1\ldots264} \overline{x}^i_2 \quad \cdots \quad \min_{i=1\ldots264} \overline{x}^i_W \right]. \quad (5)$$

From (4) and (5), the maximal distance can be calculated as follows:

$$\max d_w = \max \overline{x}_w - \min \overline{x}_w. \quad (6)$$

A wavelength-dependent MNR can be determined as the ratio of (6) and (3):

$$\mathrm{MNR}x_w = \left[\frac{\max d_1}{\max \sigma_1} \quad \frac{\max d_2}{\max \sigma_2} \quad \cdots \quad \frac{\max d_W}{\max \sigma_W} \right]. \quad (7)$$

The average of (7) is converted to decibels (dB) as follows:

$$\mathrm{MNR} = 20 \bullet \log_{10}\left(\frac{\sum_{w=1}^{W} \mathrm{MNR}x_w}{W} \right) [\mathrm{dB}]. \quad (8)$$

Based on the fact that SNR is expressed in decibels, equation (8) demonstrates similarities between MNR and SNR. According to Table 2, S1.7 sensor produces the highest MNR, while S2.2 sensor produces the lowest MNR. In the next section, MNR will be critical in analyzing and understanding the results.

4.2. Machine Learning. This section provides a step-by-step description of the data preparation. First, the samples are measured with NIR sensors, and the resulting intensities are saved as CSV files. The CSV files can be opened and read with Python, which is commonly used to analyze data and create machine learning models. The measured raw data are visualized in Table 3.

A measurement with sensor S1.4 resulted in the aforementioned table. This corresponds to a wavelength range of 1100–1350 nm. To reduce measurement noise, each sample is measured for more than 5 s, and the mean values over these measurements are calculated. The next step in the preprocessing process is to reshape the data in a way that can be utilized with Python's machine learning libraries.

Table 4 shows the intensity values for all samples measured by sensor S1.4 and represents one dataset for the machine learning model. Additionally, datasets from other sensors and a combination of all sensors are also available. A total of 5 datasets have now been prepared:

(a) Dataset from sensor S1.4

(b) Dataset from sensor S1.7

(c) Dataset from sensor S2.0

(d) Dataset from sensor S2.2

(e) Dataset from a combination of all sensors (S1.4 + S1.7 + S2.0 + S2.2)

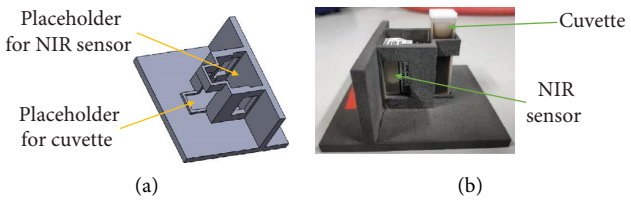


Figure 2: The design scheme of the fixture (a) and the real-life application (b).

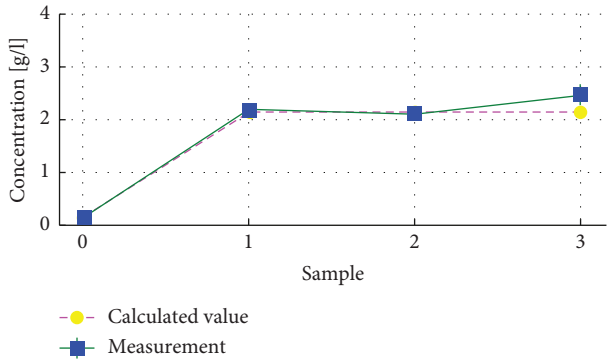
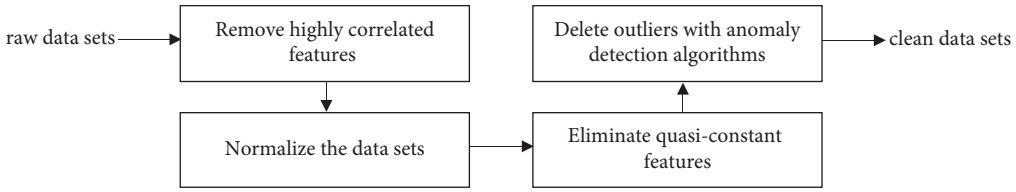


Figure 3: An initial test to determine the effect of sodium acetate on the concentration of acetic acid in samples.

Various datasets can be used to investigate which wavelength range is best suited for estimating acetic acid concentration.

The following steps were used to process these raw datasets.



A high correlation [38] of features inevitably means that they will only add dimension without adding value. To this end, similar data will be removed during the first pre-processing step since they are redundant. The normalization of the data is an essential step in machine learning [39, 40], which is especially important here since the NIR sensors have a different range of values. As a result of normalization, all data are in the range of 0 to 1. Occasionally, one measured signal remains constant throughout the process of recording data, thereby giving no additional information or value. A suggested method to overcome this problem requires the use of a signal variance [41]. The low variance signals will be categorized as constant signals and eliminated.

In the final processing step, anomaly detection is performed to eliminate outliers [42]. Most datasets contain anomalies. This could be due to problems with the sensor or disturbances from environmental conditions. After the anomaly detection step, principal component analysis (PCA) [43] was used to reduce dimension. However, the initial tests showed poorer results, so PCA was not used anymore in this study.

Then, the data are divided into training and testing data. In this paper, 20% of the data are randomly selected for testing. It is the selected 20% of the test data that determines the quality of the machine learning model. However, since the selection procedure is random, the classification accuracy may differ according to the data selected. To avoid coincidences in results, the training and test data are divided 50 times manually, and the mean and standard deviation are calculated from these 50-splitting results. An example is visualized in Table 5.

Each split generates 211 samples of training data (80%) and 53 samples of test data (20%). It is apparent from Table 5 that the samples and the order of the samples are different for each split. This method is known as cross-validation and can avoid coincidences in a result [44, 45].

As a final step, they are classified using different machine learning models [46]. In this project, the following machine learning models were used:

(a) Logistic Regression

(b) Decision Tree

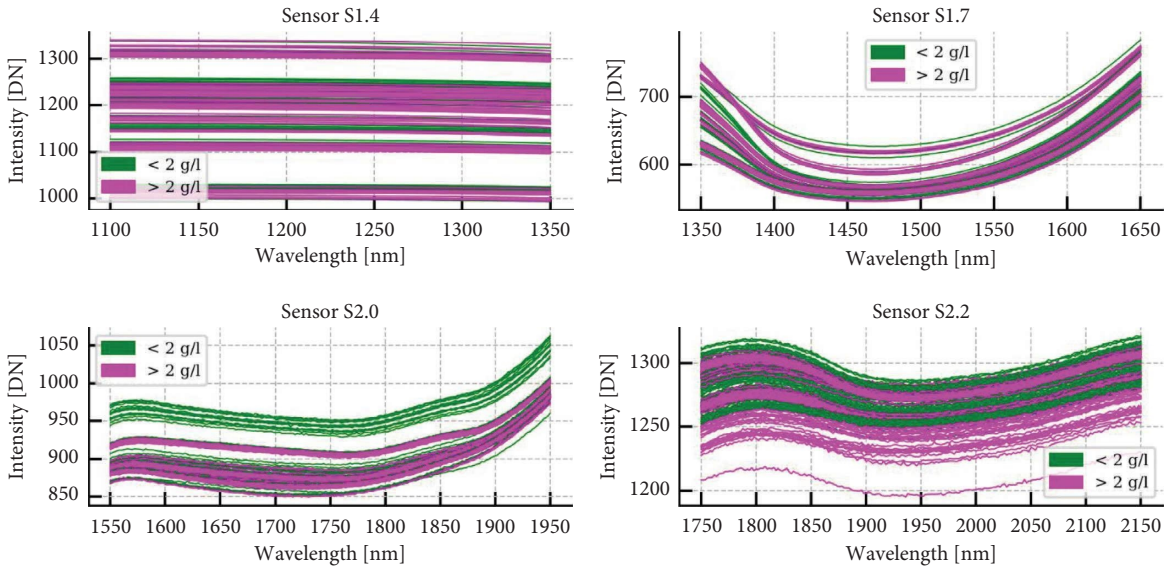


Figure 4: Diagram based on classes for a wavelength range of 1100 to 2150 nm.

Table 2: The MNR of the sensors.

NIR sensor	Wavelength range (step size = 2 (nm))	MNR (dB)
S1.4	1100–1350	34.90
S1.7	1350–1650	38.36
S2.0	1550–1950	30.18
S2.2	1750–2150	29.36

Table 3: A visual representation of raw measurement data (sample 1, sensor S1.4).

Timestamp/ wavelength	1100 nm	1102 nm	...	1348 nm	1350 nm
31.03.2022 11:13:33	1207.87	1207.82	...	1194.84	1194.10
...	...	...	...	...	...
31.03.2022 11:13:40	1208.79	1209.18	...	1195.51	1195.46

Table 4: A visual representation of the transformed data (all samples, sensor S1.4).

Sample/ wavelength, class	1100 nm	1102 nm	...	1348 nm	1350 nm	Class
1	1208.30	1208.29	...	1194.87	1194.63	0
...	...	...	...	...	...	...
264	1326.00	1325.83	...	1311.09	1310.74	1

(c) K-Nearest Neighbors (KNN)

(d) Linear Discriminant Analysis

(e) Gaussian Naïve Bayes

(f) Support Vector Machine

(g) Random Forest

The accuracy from the confusion matrix [47] and the learning curve [48] are used as evaluation criteria. The confusion matrix (Table 6) is defined as follows.

The samples in class 0 have less than 2 g/l of acetic acid, while the samples in class 1 have more than 2 g/l. If the predicted classes match the actual classes, the sample is classified as either true positive (class 1) or true negative (class 0). The accuracy from the confusion matrix (Table 6) can then be calculated using the following equation:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{9}$$

5. Results

264 samples were used to generate 5 different types of datasets, which were then combined with 7 machine learning models and 2 evaluation criteria to produce a total of 70 results. It is intended to provide a high level of clarity by showing only a few results (Table 7) in this paper. The other results to support the findings of this study are included within the supplementary information file.

The accuracy scores (9) calculated in Table 7 were based on test data. The random forest model provides the highest mean accuracy of 82.7% with a standard deviation of 5.6%. A logistic regression model is used for comparison with other more complex models and has the highest accuracy of 56% and a standard deviation of 6%. Although the K-nearest neighbors model is not a complex machine learning model, its mean accuracy is comparable to the random forest model.

The dataset from all sensors produced higher accuracy than the dataset from the individual sensors. A wide wavelength range is likely to explain this result. The standard deviation of accuracy is strongly influenced by the number of measurements. In the initial study, which is not included in this work, the standard deviation was over 10% for 90 measurements.

Table 5: An example of manual cross-validation between training and test data.

Split	Training data				Test data		
1	Sample 41	Sample 79	...	Sample 35	Sample 86	...	Sample 32
2	Sample 26	Sample 56	...	Sample 23	Sample 4	...	Sample 65
...	...	...	...	...	...	...	
50	Sample 62	Sample 42	...	Sample 31	Sample 21	...	Sample 7

Table 6: Confusion matrix.

		Actual classes	
		1	0
Predicted classes	1	True positive	False positive
	0	False negative	True negative

Table 7: A comparison of the accuracy of several machine learning models.

Dataset	Machine learning model	Accuracy (mean) (%)	Accuracy (std. deviation) (%)
All sensors	Random forest	82.7	5.6
S2.2	Logistic regression	56.1	6.0
S1.7	Logistic regression	46.9	6.6
All sensors	K-nearest neighbors	79.0	5.2

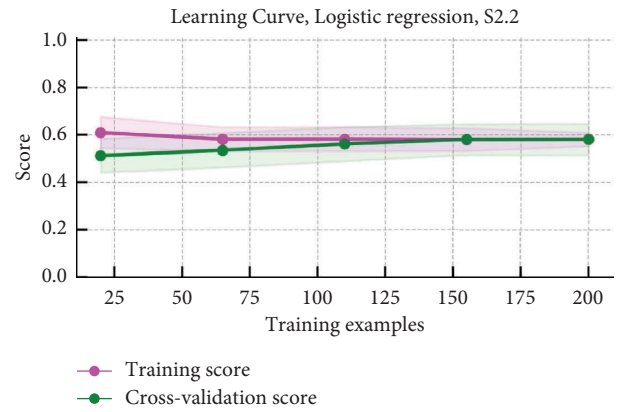


Figure 5: Learning curve for logistic regression model.

A detailed analysis of the accuracy of all 5 datasets and 7 machine learning models is included in the supplementary material (Figures S1–S35). Furthermore, other evaluation criteria such as "recall," "precision," and "F1-score" are also provided (Tables S1–S7).

Of the individual sensors, S1.7 provides the lowest accuracies and S2.2 provides the highest accuracies. This can be explained by an examination of the MNR for each of the sensors. According to Table 2, sensor S1.7 has the highest MNR of 38.36 dB and sensor S2.2 has the lowest MNR of 29.36 dB.

Low MNRs indicate that the sensor is very sensitive to the measurements, resulting in high classification accuracy. However, noise must be reduced through a large number of measurements. In contrast, S1.7 sensor has the highest MNR, making it the most robust sensor. The robustness makes it difficult to distinguish similar acetic acid concentrations between samples. According to this result, a slightly robust sensor is not necessarily profitable when used in machine learning classification.

Following are two diagrams that illustrate the learning curves of logistic regression and random forest models. "All sensors" in the title refer to all measurements taken on sensors S1.4, S1.7, S2.0, and S2.2. The plots show the number of training examples on the x-axis and the score or accuracy (0-1) on the y-axis. The solid lines indicate the calculated mean of the scores, and the shaded area shows the ±1 standard deviation region. Figure 5 shows the learning curve for the logistic regression model, and it consists of training data (magenta) and test data (green).

According to Figure 5, the two lines converge to a score of approximately 0.6. It is known as a high-bias, low-variance model. In this case, more training data will not lead to a higher convergence point/accuracy, and another machine learning model with higher complexity should be used instead. Therefore, the random forest model was used, and its learning curve is shown in Figure 6.

The complexity of a random forest model can be determined by the number of trees in the forest (n_estimator) and the maximum depth of the tree (max_depth). These parameters were set to the following values: n_estimator = 100 and max_depth = 5, to avoid overfitting and keep the complexity of the random forest model high. Figure 6 illustrates a learning curve for a model with a low bias and a slightly high variance. More training data can reduce the distance between the two lines. It is also possible to perform extensive hyperparameter optimization on machine learning models if higher accuracy is required.

As shown in other studies [49], reducing the quantity of spectral data can improve accuracy. By removing highly correlated and quasi-constant features, the amount of NIR spectral data has been reduced during the processing step in this study. As an alternative, relevant features can also be extracted manually. For example,

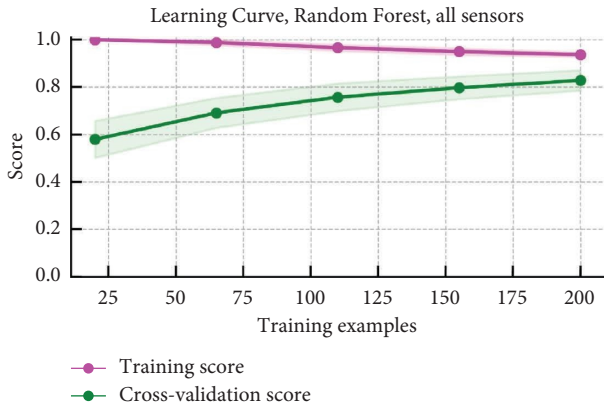


Figure 6: Learning curve for random forest model.

intensities around 1680 nm and 1724 nm can be selected as training data [50].

Moreover, UV/Vis spectroscopy can be applied instead of the NIR sensor [51, 52]. It allows comparison of the results with the NIR sensors as well as the possibility of combining both measurements as new training data for the machine learning model. Lastly, the intensities are averaged over time directly following measurements. Another possible improvement is to use a technique other than averaging, e.g., digital filtering or Fourier transformation [53].

As a next step, the laboratory's NIR measurement will be compared to the biogas plant's installed sensors. Once the machine learning results from the on-line measurement are satisfactory, the entire process in the flowchart from Section 3.1 can be spared. As a result, considerable time and resources are saved, and the actual concentration of acetic acid in the digester can be monitored live.

Furthermore, the same procedure can be executed to calibrate NIR sensors based on other acids, dry matter, and pH values. If all of these variables could be accessed in real time, then a fully automated flexible biogas production system could be achieved.

6. Conclusion

NIR sensors can be used to classify the acetic acid concentration in the biogas substrates. The dataset from all sensors produced higher accuracy than the dataset from the individual sensors. Several machine learning models were tested, with random forest producing the most accurate results. Hyperparameter optimization can improve results, but it must be done carefully to avoid overfitting.

During the data processing phase, reducing the quantity of spectral data is an integral part of improving accuracy. In this study, the dataset dimension was significantly reduced by removing highly correlated features, eliminating quasi-constant features, and deleting outliers with anomaly detection algorithms.

Various datasets should be used for training and testing the algorithms several times for ensuring that the results are not influenced by coincidence. Mean and standard deviation should always be included for all machine learning results.

A sensor's MNR can help to determine whether it can be used efficiently for classifying data using machine learning. Sensor S1.7 with the highest MNR produced on average poor accuracy while sensor S2.2 with the lowest MNR produced high accuracy.

Symbols

d: Distance between $\max \overline{x}_w$ and $\min \overline{x}_w$
N: Number of measurements at a specific wavelength during a measurement period
W: Number of wavelengths
x: Measurement
$\overline{x}$: The arithmetic average of the measurement

Greek Letters

σ: Standard deviation of the measurement

Sub and Superscripts

i: Index for samples: 1, . . ., 264
t: Index for measurements with a sampling time of 1 s
w: Index for wavelength with a resolution of 2 nm, e.g., 1102 nm and 1256 nm

Abbreviations

CHP: Combined heat and power
dB: Decibel
EEG: Renewable energy sources act
FN: False negative
FP: False positive
GC: Gas chromatography
HPLC: High-performance liquid chromatography
KNN: K-nearest neighbors
LDA: Linear discriminant analysis
MNR: Maximal distance-to-noise ratio
NIR: Near infrared
PCA: Principal component analysis
PLC: Programmable logic controller
SNR: Signal-to-noise ratio
TN: True negative
TP: True positive
UV/Vis: Ultraviolet-visible.

Data Availability

The data used to support the findings of this study are included within the supplementary information file.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Fachagentur Nachwachsende Rohstoffe (2220NR046A).

Supplementary Materials

The additional information consists of all the results from other classification metrics calculated from the confusion matrix, such as "accuracy," "recall," "precision," and "F1-score" (Tables S1–S7). In addition, learning curves for all five datasets and seven machine learning models (Figures S1–S35) are presented in the supplementary material as well. The following is a concise description of each additional material. Table S1: classification metrics for logistic regression model. Table S2: classification metrics for decision tree model. Table S3: classification metrics for K-nearest neighbors model. Table S4: classification metrics for linear discriminant analysis model. Table S5: classification metrics for Gaussian Naïve Bayes model. Table S6: Classification metrics for support vector machine model. Table S7: classification metrics for random forest model. Figure S1: learning curve for logistic regression model, sensor S1.4. Figure S2: learning curve for logistic regression model, sensor S1.7. Figure S3: learning curve for logistic regression model, sensor S2.0. Figure S4: learning curve for logistic regression model, sensor S2.2. Figure S5: learning curve for logistic regression model, all sensors. Figure S6: learning curve for decision tree model, sensor S1.4. Figure S7: learning curve for decision tree model, sensor S1.7. Figure S8: learning curve for decision tree model, sensor S2.0. Figure S9: learning curve for decision tree model, sensor S2.2. Figure S10: learning curve for decision tree model, all sensors. Figure S11: learning curve for K-nearest neighbors model, sensor S1.4. Figure S12: learning curve for K-nearest neighbors model, sensor S1.7. Figure S13: learning curve for K-nearest neighbors model, sensor S2.0. Figure S14: learning curve for K-nearest neighbors model, sensor S2.2. Figure S15: learning curve for K-nearest neighbors model, all sensors. Figure S16: learning curve for linear discriminant analysis model, sensor S1.4. Figure S17: learning curve for linear discriminant analysis model, sensor S1.7. Figure S18: learning curve for linear discriminant analysis model, sensor S2.0. Figure S19: learning curve for linear discriminant analysis model, sensor S2.2. Figure S20: learning curve for linear discriminant analysis model, all sensors. Figure S21: learning curve for Gaussian Naïve Bayes model, sensor S1.4. Figure S22: learning curve for Gaussian Naïve Bayes model, sensor S1.7. Figure S23: learning curve for Gaussian Naïve Bayes model, sensor S2.0. Figure S24: learning curve for Gaussian Naïve Bayes model, sensor S2.2. Figure S25: learning curve for Gaussian Naïve Bayes model, all sensors. Figure S26: learning curve for support vector machine model, sensor S1.4. Figure S27: learning curve for support vector machine model, sensor S1.7. Figure S28: learning curve for support vector machine model, sensor S2.0. Figure S29: learning curve for support vector machine model, sensor S2.2. Figure S30: learning curve for support vector machine model, all sensors. Figure S31: learning curve for random forest model, sensor S1.4. Figure S32: learning curve for random forest model, sensor S1.7. Figure S33: learning curve for random forest model, sensor S2.0. Figure S34: learning curve for random forest model, sensor S2.2. Figure S35: learning curve for random forest model, all sensors. (*Supplementary Materials*)

References

[1] W. Krewitt and J. Nitsch, "The German Renewable Energy Sources Act—an investment into the future pays off already today," *Renewable Energy*, vol. 28, no. 4, pp. 533–542, 2003.

[2] S. Theuerl, C. Herrmann, M. Heiermann et al., "The future agricultural biogas plant in Germany: a vision," *Energies*, vol. 12, no. 3, p. 396, 2019.

[3] Fachagentur Nachwachsende Rohstoffe e. V. (FNR), "Flexibilisierung von Biogasanlagen," in *Biogas Bedarfsgerecht Nutzen Collaborators: Schütte, Andreas, ed., Fachagentur Nachwachsende Rohstoffe e. V. (FNR), A. Öffentlichkeitsarbeit, Fachagentur Nachwachsende Rohstoffe e. V. (FNR)*, Ostbevern, Germany, 2018.

[4] V. Cmp, Edited by J. Kretzschmar, S. Weinrich, and D. Pfeiffer, Eds., in *Proceedings of the International Conference on Monitoring and Process Control of Anaerobic Digestion Processes*Leipzig, Germany, March 2021.

[5] S. Jayaraj, B. Deepanraj, and S. Velmurugan, "Study on the effect of pH on biogas production from food waste by anaerobic digestion," in *Proceedings of the 9th Annual Green Energy Conference*, pp. 25–28, Tianjin, China, 2014.

[6] Collection of Methods for Biogas, "Methods to determine parameters for analysis purposes and parameters that describe processes in the biogas sector," in *Biomass Energy Use*, J. Liebetrau and D. Pfeiffer, Eds., Deutsches Biomasseforschungszentrum gemeinnützige GmbH, Leipzig, 2nd ed edition, 2020.

[7] L. Peters, F. Uhlenhut, P. Biernacki, and S. Steinigeweg, "Status of demand-driven biogas concepts to cover residual load rises," *ChemBioEng Reviews*, vol. 5, no. 3, pp. 163–172, 2018.

[8] D. G. Mulat, H. F. Jacobi, A. Feilberg, A. P. S. Adamsen, H.-H. Richnow, and M. Nikolausz, "Changing feeding regimes to demonstrate flexible biogas production: effects on process performance, microbial community structure, and methanogenesis pathways," *Applied and Environmental Microbiology*, vol. 82, no. 2, pp. 438–449, 2016.

[9] D. J. Batstone, J. Keller, I. Angelidaki et al., "The IWA anaerobic digestion model No 1 (ADM1)," *Water Science and Technology*, vol. 45, no. 10, pp. 65–73, 2002.

[10] W. Budhijanto, C. W. Purnomo, and N. C. Siregar, "Simplified mathematical model for quantitative analysis of biogas production rate in a continuous digester," *EJH*, vol. 16, no. 5, pp. 167–176, 2012.

[11] G. Mandy and S. Roland, "An analysis of available mathematical models for anaerobic digestion of organic substances for production of biogas," in *Proceedings of the International Gas Union Research Conference*, Paris, France, 2008.

[12] T. Beltramo, M. Klocke, and B. Hitzmann, "Prediction of the biogas production using GA and ACO input features selection method for ANN model," *Information Processing in Agriculture*, vol. 6, no. 3, pp. 349–356, 2019.

[13] S. Cinar, S. O. Cinar, N. Wieczorek, I. Sohoo, and K. Kuchta, "Integration of artificial intelligence into biogas plant operation," *Processes*, vol. 9, no. 1, p. 85, 2021.

[14] Z. Wang, X. Peng, A. Xia et al., "The role of machine learning to boost the bioenergy and biofuels conversion," *Bioresource Technology*, vol. 343, Article ID 126099, 2022.

[15] E. Mauky, S. Weinrich, H.-J. Nägele, H. F. Jacobi, J. Liebetrau, and M. Nelles, "Model predictive control for demand-driven

biogas production in full scale," *Chemical Engineering and Technology*, vol. 39, no. 4, pp. 652–664, 2016.

[16] E. A. Mauky, *A Model-Based Control Concept for a Demand-Driven Biogas Production*, Springer, Berlin, Germany, 2018.

[17] L. Peters, P. Biernacki, F. Uhlenhut, and S. Steinigeweg, *The Economy, Sustainable Development, and Energy International Conference,* MDPI, Basel Switzerland, 2018.

[18] M. Harandi, J. Taheri, and B. C. Lovell, *Machine Learning Algorithms for Problem Solving in Computational Applications*, S. Kulkarni, Ed., IGI Global, Pennsylvania, USA, 2012.

[19] S. Kulkarni, Ed., *Machine Learning Algorithms for Problem Solving in Computational Applications*, IGI Global, Pennsylvania, USA, 2012.

[20] S. B. Goldberg, N. Flemotomos, V. R. Martinez et al., "Machine learning and natural language processing in psychotherapy research: alliance as example use case," *Journal of Counseling Psychology*, vol. 67, no. 4, pp. 438–448, 2020.

[21] T. Hutchinson, "Natural language processing and machine learning as practical toolsets for archival processing," *RMJ*, vol. 30, no. 2, pp. 155–174, 2020.

[22] V. T. Hayashi and W. V. Ruggiero, "Hands-free authentication for virtual assistants with trusted IoT device and machine learning," *Sensors*, vol. 22, no. 4, p. 1325, 2022.

[23] M. A. Khan, A. Tripathi, A. Dixit, and M. Dixit, "Correlative analysis and impact of intelligent virtual assistants on machine learning," in *Proceedings of the 2019 11th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Honolulu, HI, USA, January 2019.

[24] H.-N. Guo, S.-B. Wu, Y.-J. Tian, J. Zhang, and H.-T. Liu, "Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: a review," *Bioresource Technology*, vol. 319, Article ID 124114, 2021.

[25] G. H. Gu, J. Noh, I. Kim, and Y. Jung, "Machine learning for renewable energy materials," *Journal of Materials Chemistry*, vol. 7, no. 29, pp. 17096–17117, 2019.

[26] A. Morellos, X.-E. Pantazi, D. Moshou et al., "Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy," *Biosystems Engineering,* vol. 152, pp. 104–116, 2016.

[27] S. Nawar and A. M. Mouazen, "On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning," *Soil and Tillage Research*, vol. 190, pp. 120–127, 2019.

[28] S. Fuentes, E. Tongson, R. R. Unnithan, and C. Gonzalez Viejo, "Early detection of aphid infestation and insect-plant interaction assessment in wheat using a low-cost electronic nose (E-Nose), near-infrared spectroscopy and machine learning modeling," *Sensors*, vol. 21, no. 17, p. 5948, in press, 2021

[29] H. Parastar, G. van Kollenburg, Y. Weesepoel, A. van den Doel, L. Buydens, and J. Jansen, "Integration of handheld NIR and machine learning to "Measure and Monitor" chicken meat authenticity," *Food Control*, vol. 112, Article ID 107149, 2020.

[30] N. Vélez Rivera, J. Gómez-Sanchis, J. Chanona-Pérez et al., "Early detection of mechanical damage in mango using NIR hyperspectral images and machine learning," *Biosystems Engineering*, vol. 122, pp. 91–98, 2014.

[31] A. Stockl and H. Oechsner, "Near-infrared spectroscopic online monitoring of process stability in biogas plants," *Engineering in Life Science*, vol. 12, no. 3, pp. 295–305, 2012.

[32] L. C. Krapf, A. Gronauer, U. Schmidhalter, and H. Heuwinkel, "Near infrared spectroscopy calibrations for the estimation of process parameters of anaerobic digestion of energy crops and livestock residues," *Journal of Near Infrared Spectroscopy*, vol. 19, no. 6, pp. 479–493, 2011.

[33] L. C. Krapf, A. Gronauer, U. Schmidhalter, and H. Heuwinkel, "Evaluation of agricultural feedstock-robust near infrared calibrations for the estimation of process parameters in anaerobic digestion," *Journal of Near Infrared Spectroscopy*, vol. 20, no. 4, pp. 465–476, 2012.

[34] F. Graf, Ed., *Biogas: Erzeugung, Aufbereitung, Einspeisung*, Oldenbourg Industrieverl, München, 2011.

[35] F. Lichti, S. Tappen, and J. Schober, "Abschlussbericht zum vorhaben," vol. 14/13, 2018.

[36] M. Wan, M. Qu, W. Hu et al., "Estimation of soil pH using PXRF spectrometry and Vis-NIR spectroscopy for rapid environmental risk assessment of soil heavy metals," *Process Safety and Environmental Protection*, vol. 132, pp. 73–81, 2019.

[37] M. Yang, D. Xu, S. Chen, H. Li, and Z. Shi, "Evaluation of machine learning approaches to predict soil organic matter and pH using vis-NIR spectra," *Sensors*, vol. 19, no. 2, p. 263, 2019, in press.

[38] M. Pal and P. Bharati, *Applications of Regression Techniques*, M. Pal and P. Bharati, Eds., Springer Singapore, Berlin, Germany, 2019.

[39] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, August 2020.

[40] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, Article ID 105524, 2020.

[41] R. M. Gray and L. D. Davisson, *An Introduction to Statistical Signal Processing*, Cambridge University, Cambridge, UK, 2006.

[42] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[43] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[44] C. A Ramezan, T. A Warner, and A. E Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sensing*, vol. 11, no. 2, p. 185, 2019.

[45] D. Berrar, *Encyclopedia of Bioinformatics and Computational Biology,* Elsevier, Amsterdam, Netherlands, 2019.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort et al., *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] A. Kulkarni, D. Chong, and F. A. Batarseh, *Data Democracy,* Elsevier, Amsterdam, Netherlands, 2020.

[48] T. Viering and M. Loog, *The Shape of Learning Curves: A Review*, arXiv, 2021.

[49] A. G. Methrom, *NIR Spectroscopy: A Guide to Near-Infrared Spectroscopic Analysis of Industrial Manufacturing Processes*, Methrom AG, Herisau, Switzerland, 2013.

[50] H. Chung and M.-S. Ku, "Feasibility of monitoring acetic acid process using near-infrared spectroscopy," *Vibrational Spectroscopy*, vol. 31, no. 1, pp. 125–131, 2003.

[51] D. P. Mesquita, C. Quintelas, A. L. Amaral, and E. C. Ferreira, "Monitoring biological wastewater treatment processes: recent advances in spectroscopy applications," *Reviews in Environmental Science and Biotechnology*, vol. 16, no. 3, pp. 395–424, 2017.

[52] C. Wolf, D. Gaida, A. Stuhlsatz, T. Ludwig, S. McLoone, and M. Bongards, "Predicting organic acid concentration from UV/vis spectrometry measurements–a comparison of machine learning techniques," *Transactions of the Institute of Measurement and Control*, vol. 35, no. 1, pp. 5–15, 2013.

[53] U. Hassan and M. S. Anwar, "Reducing noise by repetition: introduction to signal averaging," *European Journal of Physics*, vol. 31, no. 3, pp. 453–465, 2010.