

Prediction of the Methane Production in Biogas Plants Using a Combined Gompertz and Machine Learning Model

Bolette D. Hansen[1,2(✉)], Jamshid Tamouk[2], Christian A. Tidmarsh[2], Rasmus Johansen[2], Thomas B. Moeslund[1], and David G. Jensen[2]

[1] Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark
bdha@create.aau.dk
[2] EnviDan A/S, Vejlsøvej 23, 8600 Silkeborg, Denmark

Abstract. Biogas production is a complicated process and mathematical modeling of the process is essential in order to plan the management of the plants. Gompertz models can predict the biogas production, but in co-digestion, where many feedstocks are used it can be hard to obtain a sufficient calibration, and often more research is required in order to find the exact calibration parameters. The scope of this article is to investigate if machine learning approaches can be used to optimize the predictions of Gompertz models. Increasing the precision of the models is important in order to get an optimal usage of the resources and thereby ensure a more sustainable energy production. Three models were tested: A Gompertz model (Mean Absolute Percentage Error (MAPE) = 9.61%), a machine learning model (MAPE = 4.84%), and a hybrid model (MAPE = 4.52%). The results showed that the hybrid model could decrease the error in the predictions with 53% when predicting the methane production one day ahead. When encountering an offset in the predictions the reduction of the error was increased to 66%.

Keywords: Biogas · Prediction · Gompertz model · Machine learning

1 Introduction

Climate changes and increasing energy demands have increased the focus on renewable energy in the recent years. The biogas industry contributes to this by e.g., producing energy from wastewater and reducing pollution from agriculture. The biogas industry has had a significantly progress in Europe in the recent years, where the capacity was almost tripled in the period from 2007 to 2017 [1].

Biogas is produced by anaerobic digestion of organic materials such as urban waste for example food, sludge and garden waste, animal manure, industrial waste, lignocellulosic materials such as various types of straw and the biomass of microalgae [2].

How fast a feedstock can be transformed to methane depends on the composition of the feedstock. Carbohydrates, protein and fat are easy digestible, whereas e.g. lignocellulosic materials are much harder to digest [2].

In addition to this, several parameters influence the process. These are, among others, the pH, carbon/nutrient rate, organic loading rate, temperature, the microorganisms

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 734–745, 2020.
https://doi.org/10.1007/978-3-030-58799-4_53

and enzymes present, presence of some types of fatty acids, and usage of substrates. Over or under representation of some of the parameters can even lead to failure of the system [3, 4].

Mono-digestion of some feedstocks can be problematic, as it might bring the system out of balance, for instance changing the carbon/nutrient rate will lead to a too high concentration of problematic fatty acids or change the pH. Therefore, much research has focused on co-digestion of feedstocks, which can lead to an increase in the biogas production of 25–400% [3]. However, co-digestion complicates the digestion process further and therefore, mathematical modelling of the process is essential in order to keep the balance and avoid failure [3]. In 2002 [5] published the IWA Anaerobic Digestion Model No.1 (ADM1). The ADM1 can simulate an average trend in the different parameters, however, it cannot simulate the immediate variations [6]. Since the ADM1 was published plugins with modifications have been added and in 2015 [7] states that despite new models have been developed, there is readily justification for developing a new ADM2. In order to develop an ADM2 it is necessary to have a uniform approach to the mechanisms and challenges which will require further research. After 2015, other models and optimizations of the ADM1 have been performed [8].

Where the ADM1 model focuses on modelling the whole process, another model type, called Gompertz models, is often used for prediction of the gas production. This type of model can give very accurate predictions of the methane production [9], however, like the ADM1 model, this model requires a precise calibration. In order to calibrate a Gompertz model, Gompertz functions need to be set up for all the biomasses used in the model. Several studies have been made in order to improve Gompertz models by experimentally finding the kinematic parameters describing the methane production for a wide range of feedstocks [9–11]. For this reason, the literature can provide the kinematic parameters for a large variety of feedstocks. If the Gompertz functions for some of the feedstocks in a biogas plant is not available in the literature, they can be found experimentally. However, as this often takes several months, they are often based on expert knowledge instead.

It is worth noticing that despite Gompertz variables are available for one feedstock, there might be local variations in the composition of the feedstocks influencing the gas production. This is complicated further from co-digestion and other parameters influencing the digestion of each material. For this reason and despite making several tests in the laboratory, the results might not fit well in a real case scenario.

Despite that the mathematical models are essential in order to ensure a stable process, and avoid failure, they are very hard to calibrate, especially when the parameter complexity is high. In these cases further research in parameter characterization is required [3].

Machine learning models have been developed in order to predict and optimize the methane production. In wastewater treatment plants Neural Networks have for instance been used to find the optimal settings for as high a methane yield as possible [12]. In agricultural biogas plants a combination of Genetic Algorithm and Ant Colony Optimization has been used to predict the present gas production based on measured process variables [13]. In controlled laboratory scale experiments Neural Networks have shown precise predictions of the biogas production [14]. Likewise, the machine

learning algorithms Random Forest and XGBoost have shown efficient future predictions of the methane yield in an industrial-scale co-digestion facility [15].

Being able to predict the future biogas production is essential in order to plan the operation of the biogas plant. In some cases, the goal is to produce as much methane as possible, as there is an unmet demand. In other cases, it is essential to keep as stable a production as possible in order to meet the demand while avoiding overproduction. In case of overproduction, the process can be artificially inhibited by chemicals or the surplus methane can be burned off. In both cases resources are not optimally used. This has led to development of software tools used for planning of the biogas production. The software has a framework to take in different machine learning models for prediction of the biogas production. In this case the machine learning models were used to predict the production in three categories: Low, medium, and high [16].

The scope of this article is to investigate if machine learning approaches can be used to optimize the predictions of Gompertz models in industrial settings. This is done by comparison of a Gompertz model, a machine learning model, and a combined model.

2 Method

2.1 Biogas Plant

The biogas plant has a capacity at approximately 220,000 t biomass/year and produces almost 10 mio. Normal cubic meter (Nm^3) methane/year, corresponding to 99,7 GWh heat per year. The main feedstocks used in the plant are seaweed, manure, eulat and pectin. However, in total 18 specific feedstocks were used. It is worth noticing that it is an industrial setting and the available feedstocks changes over time. For this reason, some of the feedstocks were only used in the first half of the measurement period, while others were only used in the second half of the period. Fortunately, the amount of these temporary feedstocks is limited, and the models needs to be tolerant to this type of changes.

In this biogas plant a Gompertz model is used to plan the infeed to the plant in order to obtain as constant a biogas production as possible.

2.2 Data Set

The data set obtained from the biogas plant contained consecutive measures from 818 days. In total 18 different feedstocks were used of which one was only used in the first half of the data set and four were only used in the last half of the data set.

2.3 Gompertz Model

For each biomass the expected methane production over time can be described by Gompertz functions [17, 18]. The formula for Gompertz functions can be seen in Eq. 1.

$$P(t) = P_{max} \cdot \exp\left(-\exp\left(\frac{R_{max} \cdot e}{P_{max}} \cdot (\lambda - t) + 1\right)\right) \quad (1)$$

Where P is the culminative methane production, P_{max} is the maximal total methane production R_{max} is the maximal methane production rate, t is the time measured in days and λ is the delay before any gas is produced from the specific biomass. Hereafter the methane production for a biomass on a given day can be calculated as seen in Eq. 2.

$$P_{day} = P(t_{day}) - P(t_{day} - 1) \quad (2)$$

Where P_{day} is the amount of produced methane on the specific day and t_{day} is the day for prediction. The expected daily methane production per ton of each biomass can be seen in Fig. 1.

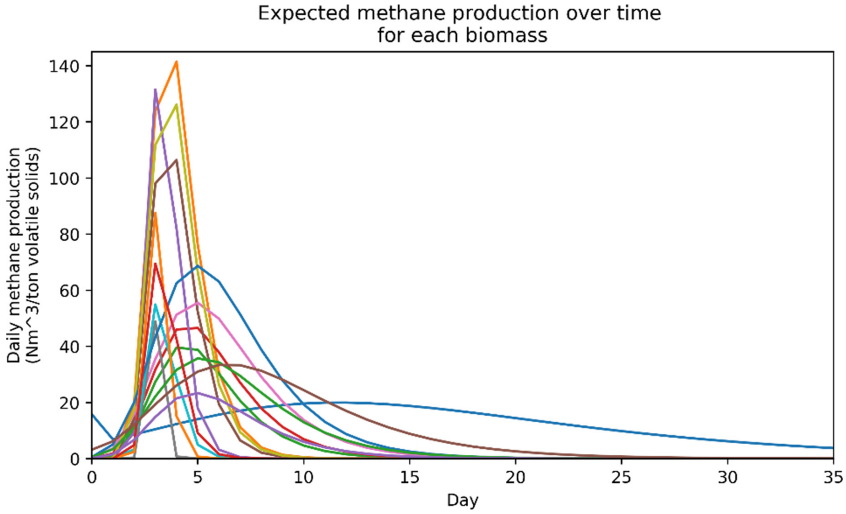


Fig. 1. Expected methane produced by each type of biomass over time

The methane production was then calculated as the sum of the added methane per biomass per day, including biomasses added up to 60 previous to the precent day. The model encountered the retention as a percentage removal of biomass per day.

The parameters in the Gompertz functions were initially based on empirical numbers where these were available. For the biomasses where empirical data was not available, they were estimated based on knowledge about similar biomasses. Hereafter the model was calibrated using the previous 90 days as calibration data. The calibration was done by minimizing the root mean square error.

In addition to the initial parameters each biomass had a span they were forced to stay within.

2.4 Machine Learning Model

Preprocessing

Before training the machine learning model, all the data points were normalized by subtracting the mean value and scaling to unit value using the Python library scikit-learn [19]. Hereafter zeros were replaced with values close to zero. This was done in order to account for algorithms being sensitive to zeros and as it showed better results in some cases. As biogas production is a time-consuming process which takes several days, eight additional features were added to each data point. These features were the measured gas production from the previous six days and the mean and standard divination calculated for the present and previous nine days. As the feature creation in this case implies inclusion of the input parameters nine days before the measurement, the first nine days of the measurement data was excluded due to missing data.

As the last part of the data set was collected while the model was developed, only the first 430 datapoints were used for training, while the last 350 datapoints were used for testing. Thereby the training set contains one biomass which is not present in the test set and the test set contains four parameters which are not present in the training data. An overview of the preprocessing pipeline can be seen in Fig. 2.

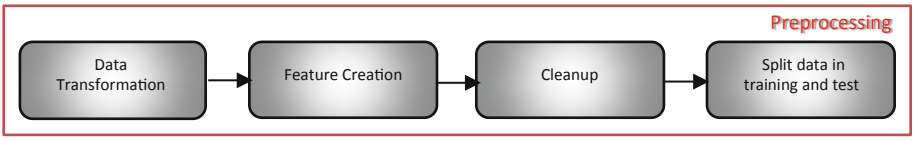


Fig. 2. The preprocessing pipeline.

Model Development and Training

The training set was split into 25 folds for folding where 23 folds were used for training, one fold was used for validation and one fold was used for test.

Initially 15 commonly used machine learning algorithms from the Python library scikit-learn [19] were tested. These were uniform k-nearest neighbors (kNN), distance kNN, Bagging [20] with Decision Tree, AdaBoost [21] Regressor with Decision Tree, Random Forest Regression [22], Bagging with Random Forest, recursive feature elimination (RFE) with the core of linear Ridge, Recursive Feature Elimination by using Gradient Boosting [23], Principal Components Regression, Quadratic Discriminant Analysis, Lasso Regression, Multilayer Perceptron (MLP) Regressor, Naive Bayes [24], Extra Trees [25], and Support Vector Machine [26]. Based on this initial test seven methods were selected. These algorithms were uniform kNN, distance kNN, recursive feature elimination (RFE) with the core of linear Ridge, MLP Regressor, Ada Boost Regressor with Decision Tree Regressor, RFE with the core of Gradient Boosting Regressor, and Random Forest Regressor.

For each fold, the best three models were ensembled whereby the prediction would be the mean of the three models. If the ensemble score was better than the score of the single models this was selected. Otherwise the best single model was selected. If the

selected model did not obtain a sufficient precision, no model from that fold was saved. Lastly all the saved models were applied to the test set and the prediction of the models were averaged in order to predict the methane production one day ahead. An overview of the setup can be seen in Fig. 3.

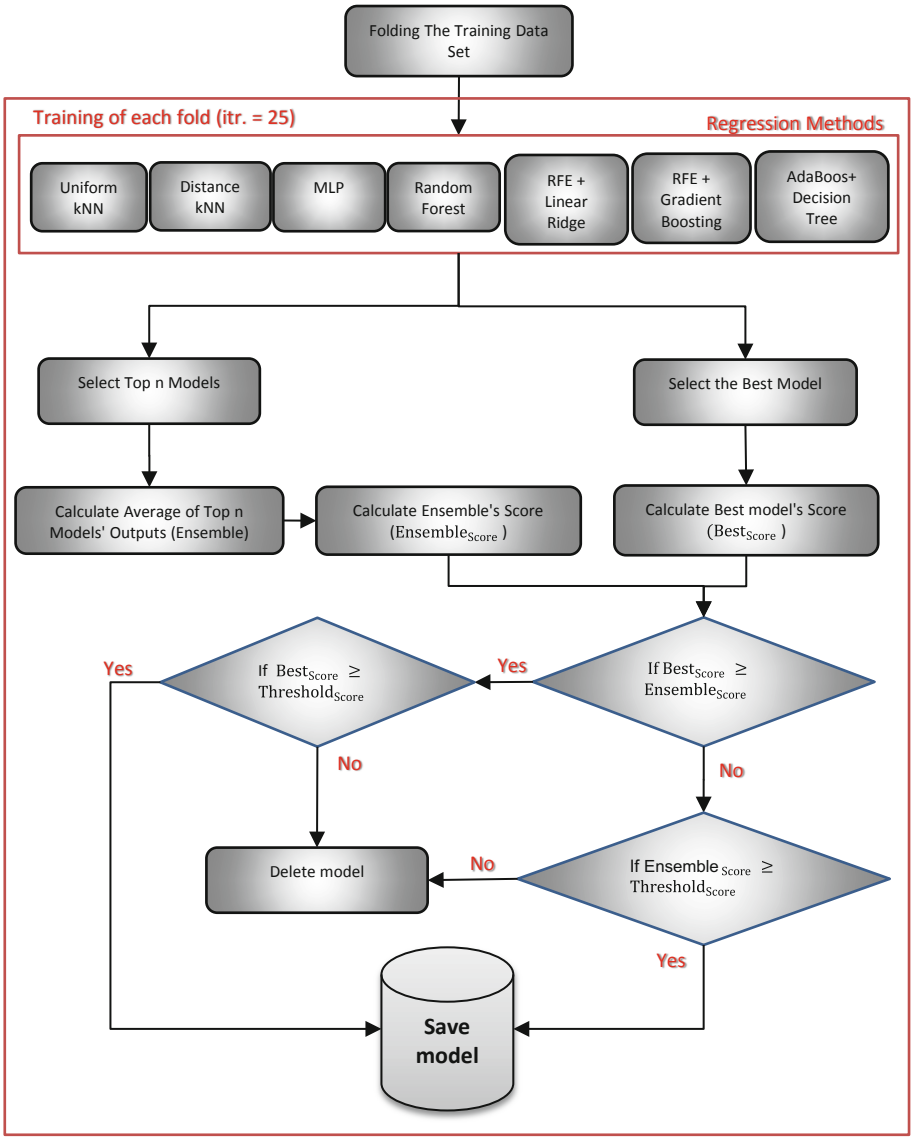


Fig. 3. Flowchart over the method for developing the machine learning model.

2.5 Hybrid Model Based on the Gompertz Model and the Machine Learning Model

In order to make a hybrid model based, the error of the Gompertz model was found by subtracting the Gompertz predictions from the measurements. Hereafter a machine learning model similar to the model described in Sect. 2.4 was trained to predict the error of the Gompertz model. Subsequently the predictions from the machine learning model were added to the predictions from the Gompertz model in order to predict the amount of methane produced.

3 Results

The predictions from respectively the Gompertz model, the machine learning model, and the hybrid model for the full test set can be seen in Fig. 4, and a zoomed version can be seen in Fig. 5. Likewise, the correlation between the observations and the predictions was found as seen in Fig. 6.

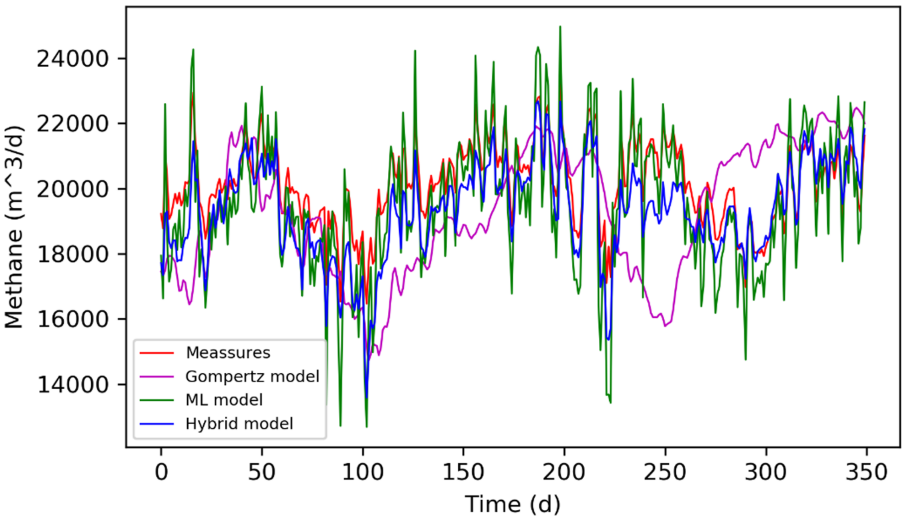


Fig. 4. Comparison of the measured biogas production and the production predicted one day ahead by the Gompertz model, the machine learning model and the hybrid model.

From Fig. 6 it was observed that there was an offset in the predictions according to the correlation line. A similar offset was observed for prediction on the training set. If calculating the mean of the measurements and the predictions and adjusting for this by adding the difference in mean values to the predictions the error of the predictions is decreased. The Mean Absolute Percentage Error (MAPE) for each of the models with and without adjustment according to the mean prediction can be seen in Table 1.

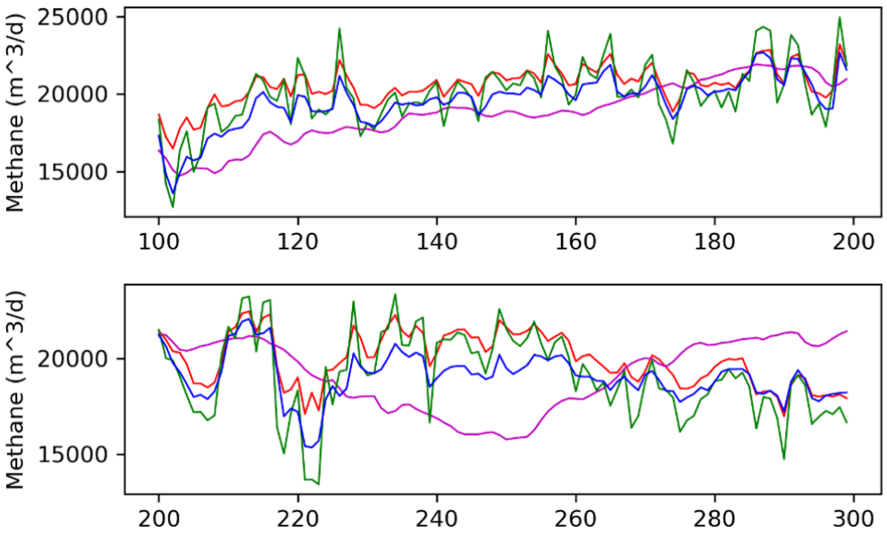


Fig. 5. Zoomed versions of Fig. 3

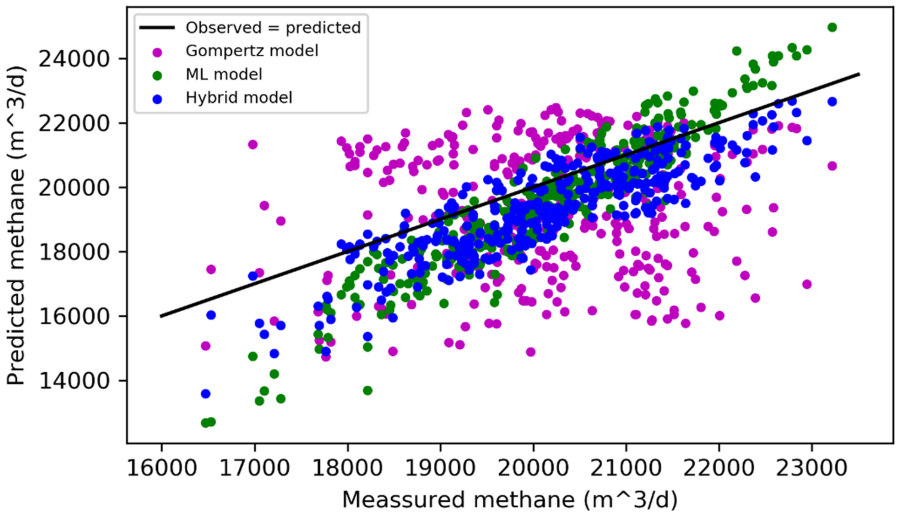


Fig. 6. Comparison between the observed and predicted values for the three models.

Table 1. The mean absolute error for each of the three models

Model	R^2	↓MAPE	↓MAPE after adjustment
Gompertz model	−0.27	9.61%	9.00%
ML model	0.72	4.84%	3.78%
Hybrid model	0.75	**4.52%**	**3.06%**

4 Discussion

Modelling of biogas plants is essential in order to keep an optimal operation of the plants. Gompertz models can be used to plan the infeed to biogas plants in order to keep a constant output. However, these models can be hard to calibrate, especially when the number of feedstocks is high. The Gompertz model presented in this paper can estimate the methane production with a MAPE at 9.61%. From Fig. 6 it can be seen that the correlation between the predicted and observed methane production for the Gompertz model is not very corelated in the data available. This is because the model is used to plan the infeed to the biogas plant in order to ensure a constant production and it tells us that the model is used to its limits. Due to the complexity in co-digestion scenarios more research is required in order to calibrate the Gompertz model further. As this would require several tests each lasting for several days a machine learning and a hybrid model were proposed in order to optimize the predictions further.

From Figs. 4, 6 and Table 1 it can be seen that the machine learning and combined model can improve the prediction one day ahead despite the usage of additional feedstocks in the test set. When comparing the machine learning model with the measurements it is clear that it is able to find the relationships between the input feedstocks and the methane production. However, it typically overestimates the changes in the production: When the production increases the model predicts the increase to be bigger than it is and when the production decreases the model predicts the reduction in the production to be bigger than it is. The hybrid model does not have this problem as the Gompertz model generates a baseline for the production. In other industries it have been found beneficial to develop hybrid models as it can make the model more generalizable and requires less training data [27].

The issue with changes of input data is a typical issue in industrial cases. In this case the quantity of the additional biomasses was relatively low, but if the amount of these feedstocks was increased it could either be added to other feedstocks with similar compositions or the model could be retrained. However, retraining would require several datapoints with usage of the new feedstock.

As it appears from the results the hybrid model can optimize the predictions with 53% when not encountering the offset and with 66% when encountering an offset in the predictions. This is important as surplus methane will be burned off or the gas production will be inhibited with chemistry, as it is too expensive to build storage facilities. Likewise, if not enough methane is produced the demand is not met. Despite several studies deals with optimization of Gompertz functions in order to develop more precise Gompertz models, they are not suitable for comparison with our model. This is because they are based on experimental studies. The contribution of this paper is to show that we can increase the predictions of the methane production in an industrial setting were some of the parameters are based on expert knowledge as the exact parameters for each of the Gompertz functions are not available.

Two other articles focusing on machine learning based predictions of the methane production in industrial scale biogas plants were found [15, 16]. When compared to [16] the predictions from the machine learning model and the hybrid model presented

in this article are quite accurate. [16] replaced the numeric values for the biogas production with values of 0, 1 and 2 for low, medium and high production respectively and obtain an accuracy at 87%. However, due to the low resolution in the output prediction it is much easier to obtain a high accuracy. In our case, the methane production fluctuates between 16,000 Nm^2 and 22,000 Nm^2, which corresponds to one of the categories in [16]. [15] used Random Forest to predict the methane production for time horizons between one and 40 days and obtained R^2 values between 0.88 and 0.82. As the Gompertz model used in our study is used to plan the infeed, in order to obtain as constant a biogas yield as possible, it would not be fair to use R^2 for comparison. This is because optimal planning entails a lower R^2.

5 Conclusion

In this work we have shown that combining a Gompertz model with a machine learning model can optimize the prediction of the methane production one day ahead with up to 66% according to using a Gompertz model alone. This is important as prediction of the methane production is essential in order to keep a constant production of methane, and thereby ensure that the demand is met while avoiding overproduction. If the demand is not met the costumers will have to go elsewhere, which could lead to usage of none sustainable energy sources. If too much methane is produced the surplus will either be burned off or chemistry will be added in order to inhibit the production.

References

1. Xue, S., et al.: A systematic comparison of biogas development and related policies between China and Europe and corresponding insights. Renew. Sustain. Energy Rev. **117**, 109474 (2020). https://doi.org/10.1016/j.rser.2019.109474
2. Colla, L.M., et al.: Waste biomass and blended bioresources in biogas production. In: Treichel, H., Fongaro, G. (eds.) Improving Biogas Production. BBT, vol. 9, pp. 1–23. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10516-7_1
3. Hagos, K., Zong, J., Li, D., Liu, C., Lu, X.: Anaerobic co-digestion process for biogas production: progress, challenges and perspectives. Renew. Sustain. Energy Rev. **76**, 1485–1496 (2017). https://doi.org/10.1016/j.rser.2016.11.184
4. Scapini, T., et al.: Enzyme-mediated enhanced biogas yield. In: Treichel, H., Fongaro, G. (eds.) Improving Biogas Production. BBT, vol. 9, pp. 45–68. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10516-7_3
5. Batstone, D.J., et al.: The IWA anaerobic digestion model no 1 (ADM1). Water Sci. Technol. **45**, 65–73 (2002)
6. Derbal, K., Bencheikh-lehocine, M., Cecchi, F., Meniai, A.-H., Pavan, P.: Application of the IWA ADM1 model to simulate anaerobic co-digestion of organic waste with waste activated sludge in mesophilic condition. Biores. Technol. **100**, 1539–1543 (2009). https://doi.org/10.1016/j.biortech.2008.07.064

7. Batstone, D.J., Puyol, D., Flores-Alsina, X., Rodríguez, J.: Mathematical modelling of anaerobic digestion processes: applications and future needs. Rev. Environ. Sci. Biotechnol. **14**, 595–613 (2015). https://doi.org/10.1007/s11157-015-9376-4
8. Arzate, J.A., et al.: Anaerobic digestion model (AM2) for the description of biogas processes at dynamic feedstock loading rates. Chem. Ing. Tec. **89**, 686–695 (2017). https://doi.org/10.1002/cite.201600176
9. Ripoll, V., Agabo-García, C., Perez, M., Solera, R.: Improvement of biomethane potential of sewage sludge anaerobic co-digestion by addition of "sherry-wine" distillery wastewater. J. Clean. Prod. **251**, 119667 (2020). https://doi.org/10.1016/j.jclepro.2019.119667
10. dos Santos, L.A., et al.: Methane generation potential through anaerobic digestion of fruit waste. J. Clean. Prod. **256**, 120389 (2020). https://doi.org/10.1016/j.jclepro.2020.120389
11. Hernández-Fydrych, V.C., Benítez-Olivares, G., Meraz-Rodríguez, M.A., Salazar-Peláez, M.L., Fajardo-Ortiz, M.C.: Methane production kinetics of pretreated slaughterhouse wastewater. Biomass Bioenerg. **130**, 105385 (2019). https://doi.org/10.1016/j.biombioe.2019.105385
12. Akbaş, H., Bilgen, B., Turhan, A.M.: An integrated prediction and optimization model of biogas production system at a wastewater treatment facility. Biores. Technol. **196**, 566–576 (2015). https://doi.org/10.1016/j.biortech.2015.08.017
13. Beltramo, T., Klocke, M., Hitzmann, B.: Prediction of the biogas production using GA and ACO input features selection method for ANN model. Inf. Process. Agric. **6**, 349–356 (2019). https://doi.org/10.1016/j.inpa.2019.01.002
14. Tufaner, F., Demirci, Y.: Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models. Clean Techn. Environ. Policy **22**, 713–724 (2020). https://doi.org/10.1007/s10098-020-01816-z
15. De Clercq, D., Wen, Z., Fei, F., Caicedo, L., Yuan, K., Shang, R.: Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Sci. Total Environ. **712**, 134574 (2020). https://doi.org/10.1016/j.scitotenv.2019.134574
16. De Clercq, D., et al.: Machine learning powered software for accurate prediction of biogas production: a case study on industrial-scale Chinese production data. J. Clean. Prod. **218**, 390–399 (2019). https://doi.org/10.1016/j.jclepro.2019.01.031
17. Gompertz, B.: XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. & c. Phil. Trans. R. Soc. Lon. **115**, 513–583 (1825). https://doi.org/10.1098/rstl.1825.0026
18. Zwietering, M.H., Jongenburger, I., Rombouts, F.M., van't Riet, K.: Modeling of the bacterial growth curve. Appl. Environ. Microbiol. **56**, 1875–1881 (1990)
19. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **2011**, 2825–2830 (2011)
20. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996). https://doi.org/10.1007/bf00058655
21. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997). https://doi.org/10.1006/jcss.1997.1504
22. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/a:1010933404324
23. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**, 367–378 (2002). https://doi.org/10.1016/s0167-9473(01)00065-2
24. Zhang, H.: The optimality of Naive Bayes (2004). https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

25. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
26. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011). https://doi.org/10.1145/1961189.1961199
27. Kloss, A., Schaal, S., Bohg, J.: Combining learned and analytical models for predicting action effects. arXiv:1710.04102[cs] (2018)