Bioresource Technology 392 (2024) 130017

Contents lists available at ScienceDirect

Bioresource Technology

journal homepage: www.elsevier.com/locate/biortech

Real-time operation of municipal anaerobic digestion using an ensemble data mining framework

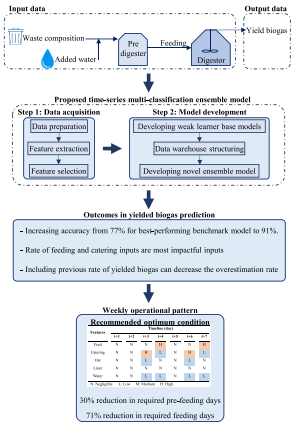
Farzad Piadeh [a,b], Ikechukwu Offie [a], Kourosh Behzadian [a,c,*], Angela Bywater [d], Luiza C. Campos [c]

[a] School of Computing and Engineering, University of West London, London W5 5RF, United Kingdom
[b] School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
[c] Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E6BT, United Kingdom
[d] Water and Environmental Engineering Group, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, SO17 1BJ, UK

HIGHLIGHTS

- Time-series ensemble model is proposed for real-time anaerobic digestion operation.
- Simple/practical features i.e. waste composition, water and feeding volume are used.
- Prediction accuracy is improved from 75% to 91% in comparison to benchmark models.
- Proposed weekly operation could reduce 70% of required feeding day operation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Anaerobic digestion
Biogas generation
Data mining
Ensemble modelling
Organic waste
Real-time operation

ABSTRACT

This study presents a novel approach for real-time operation of anaerobic digestion using an ensemble decision-making framework composed of weak learner data mining models. The framework utilises simple but practical features such as waste composition, added water and feeding volume to predict biogas yield and to generate an optimised weekly operation pattern to maximise biogas production and minimise operational costs. The effectiveness of this framework is validated through a real-world case study conducted in the UK. Comparative analysis with benchmark models demonstrates a significant improvement in prediction accuracy, increasing from the range of 50–80% with benchmark models to 91% with the proposed framework. The results also show the efficacy of the weekly operation pattern, which leads to a substantial 78% increase in biogas generation during the testing period. Moreover, the pattern contributes to a reduction of 71% in total days required for feeding and 30% in total days required for pre-feeding.

* Corresponding author.
E-mail address: kourosh.behzadian@uwl.ac.uk (K. Behzadian).

https://doi.org/10.1016/j.biortech.2023.130017
Received 23 August 2023; Received in revised form 5 November 2023; Accepted 10 November 2023
Available online 13 November 2023

1. Introduction

The exponential growth in organic waste production, primarily driven by population expansion and economic development, has become a significant global concern. This issue is accompanied by several detrimental effects, including methane emissions, water contamination, pest infestation, land degradation, escalating costs for waste management authorities, heightened health risks and immigration challenges (Awasthi et al., 2022). Over time, various waste management systems such as landfilling, open dumping, composting and incineration have been introduced (Masalegooyan et al., 2022). However, anaerobic digestion (AD) has recently gained substantial attention from both the scientific and industrial communities due to its effectiveness in organic waste management, with its production of liquid/solid fertilisers and the generation of renewable energy in the form of biogas (Gupta et al., 2023).

Accurate prediction of AD biogas production holds significant importance for several key reasons. Firstly, it enables effective energy planning by providing decision-makers with the means to estimate the potential energy output. This information is critical for assessing the feasibility and profitability of implementing these systems on a larger scale (Wang et al., 2020). Furthermore, if operators had a comprehensive understanding of the mechanisms behind optimal biogas production, they could adjust key operational variables, thus ensuring efficient utilisation of feedstock, enhancing process stability and minimising the risk of system failures (Khan et al., 2023).

Several mechanistic models have been created to facilitate the design and optimisation of AD processes in which AQUASIM, and Anaerobic Digestion Model no.1 stand out as the most rigorous and accurate models (Emebu et al., 2022). These models incorporate the conservation of mass and energy to forecast the cumulative biogas yield and its composition over time. However, despite their accuracy, the practical implementation of these models for real-time biogas prediction, capturing the intricacies and dynamic nature of AD systems and AD operation is highly challenging due to their complexity, computationally intensive nature, extensive parameters for calibration and the knowledge necessary to utilise them (Cruz et al., 2021). Alternatively, artificial intelligence (AI) models can effectively learn from the data, identify patterns and make accurate predictions or optimal decisions (Kazemi et al., 2021). The application of these models has garnered substantial interest over the past two decades, particularly in the context of optimisation and hybrid applications (Gupta et al., 2023), but many of these AI models have primarily been tested and validated at laboratory scale, limiting their applicability to industrial-scale and hindering their widespread deployment (Jia et al., 2022).

The application of AD technologies has undergone extensive testing for sludge treatment in water and wastewater treatment plants, as well as for biogas production from agricultural and livestock waste (Cruz et al., 2022) and practical implementation of AI models in real AD systems for organic municipal waste is a relatively recent development (Offie et al., 2023). Several notable research studies have employed four primary approaches: (1) machine learning models, such as feed-forward neural networks coupled with back-propagation or elastic net (Almomani, 2020; Clercq et al., 2020); (2) recurrent neural network models, specifically the nonlinear autoregressive network with exogenous inputs model (Offie et al., 2022); (3) weak learner data mining (WLDM) techniques, particularly *k*-nearest neighbourhood (KNN), Gaussian process regression (GPR), support vector machine (SVM), decision tree (DT), multiple linear regression, polynomial regression, kernel ridge regression and extreme learning machine (Wang et al., 2023; Yildirim and Ozkaya, 2023); and (4) ensemble models such as extreme gradient boosting (XGBoost) and random forest (RF) (Xu et al., 2021; Sonwai et al., 2023).

These models are typically used in numerical applications where input decision variables and predicted biogas as outputs are represented in actual volumes or concentrations. However, there is a growing demand to expand the application of AI models in the field of classification, particularly as an initial step before using these advanced models. Compared to other models, these models may offer a more straightforward approach that can allow operators to easily understand and interpret the input variables and the corresponding output classes. This simplicity of operation makes them more accessible and user-friendly for operators who may not have extensive technical expertise in advanced modelling techniques (Wang et al., 2022). Furthermore, dealing with different volumes and numbers in a practical setting can be highly challenging and cumbersome, whereas these models may simplify the decision-making process by providing clear indications of the system's operational state or the class to which inputs or outputs belongs (Yan et al., 2021).

Additionally, the relevant research models rely heavily on operational decision variables, including chemical oxygen demand, carbon-to-nitrogen ratio, ammonium concentration, temperature, organic loading rate, substrate-to-inoculum ratio, retention time, total or volatile solids, appropriate pretreatment, pH and heating condition (Fajobi et al., 2022; Sappl et al., 2023). These variables require either online monitoring of multiple parameters in field-scale applications or extensive laboratory analysis of numerous samples. However, due to technical and economic limitations often faced by AD projects, conducting such comprehensive monitoring can be challenging (Jia et al., 2022). Moreover, in practical scenarios where AD systems receive contracted daily organic waste deliveries, significantly modifying the amount of feedstock received is not a viable option. This adds another layer of complexity to the development, training, testing and operation of real-time AI models that can effectively utilise in-field data which, by its nature, suffers from technical limitations and economic considerations (Offie et al., 2023).

To the best of authors' knowledge, there has been no such a study in the past that has included the above factors for developing AI-based decision making for optimal operation of AD in real-time. Hence this study focuses on the development of a novel framework for the real-time operation of AD systems in which proposed time-series ensemble data mining framework are introduced and an optimal weekly operation pattern is provided by using only accessible operational data. The proposed framework exhibits several key innovations, making it highly suitable for real-time industrial operations. One crucial innovation lies in its simplicity and practicality, enabling straightforward application in various industrial contexts. Additionally, the incorporation of time-series concepts into the ensemble model represents another significant advancement, enhancing the model's predictive capabilities. Moreover, this framework introduces a user-friendly weekly operation pattern for easy implementation by operators that can be applicable for other worldwide projects. This feature streamlines the operational process, fostering efficient and effective utilisation of AD systems. Further details of the methodology are described in the next section, followed by the results and discussion for its application on a real-world case study of a micro-AD in London.

2. Methodology

The proposed framework for the development of an ensemble model for real-time AD operation to produce maximum biogas involves three key stages, as illustrated in Fig. 1. These stages comprise Step #1: data acquisition; Step #2: model development; and Step #3 performance assessment. The proposed methodology is explained generally such that it allows for its broad adoption in similar projects. To enhance comprehension, the methodology is demonstrated through a detailed case study explained in section 2.1, providing a clear and practical illustration of its effectiveness. Relevant operational information is collected from real-time sensors or reported by operators. The data consists of simple parameters and includes details such as the composition of local waste, including soaked oats, soaked caddy liners, tea or coffee residues, which are fed to the pre-digester after separation and

screening. Furthermore, it includes the volume of water added to the pre-digester, feeding rates to the main digester and the amount of biogas produced. The data is reported at varying intervals, ranging from a few minutes to daily. Missing data and data cleaning procedures are handled according to recommendations from Offie et al. (2023). Finally, the numerical and time-series data are transformed into features and selected based on further explanation provided in Section 2.1.

While many operational inputs, including the carbon-to-nitrogen ratio, ammonium concentration, temperature, organic loading rate, retention time, total solids, appropriate pretreatment, pH, and heating conditions can also be utilised in the development of data-driven models, this study primarily focuses on inputs related to feeding, added water and waste composition. This limitation is due this reason that many of these parameters are not consistently monitored and reported in real-time management, particularly in micro-AD projects. Note that the main feature of a given waste or biomass feedstock to be used to produce biogas is based on biochemical composition which may be similar for some types of feedstocks. However, this is not always the case as gradients constituting a type of feedstock may vary from one place to another. Hence, developing worldwide applicability in micro-AD projects can be challenging as access to biochemical composition of the feedstocks may be hard.

The selected features are employed to develop weak learner data mining (WLDM) models. These WLDM models, along with their key performance indicators (KPIs), are then stored in a data warehouse, which serves as the foundation for constructing the proposed ensemble model. Additional insights and comprehensive information on this development process are expounded in Section 2.2. Following the construction of the ensemble model, rigorous testing is conducted on real-time unseen data to evaluate its performance under real-world conditions. The outcomes of this testing and a detailed analysis of the results are discussed in Section 2.3. This analysis provides a comprehensive understanding of the model's effectiveness and its potential for real-time optimisation in practical scenarios.

2.1. Step 1: Data acquisition

Several key input parameters were considered for the analysis: (1) daily feed into the main digester tank, (2–3) added water and feed composition with various materials into the pre-digester tank, and (4) yielded biogas over the last days. Each constituent of the feed is categorised as an individual feature. To achieve this, a preliminary assessment is conducted, encompassing cross-correlation analyses between each waste category and the biogas generation, an evaluation of data accessibility, and an assessment of data reliability (refer to the work by Offie et al. (2023) for further details). While the presented framework holds applicability beyond the specific context and can be extrapolated to similar projects, its explication is based on a real case study to facilitate enhanced comprehension.

The data in this study was collected from a micro-AD plant situated in Camley Street Natural Park, Central London, UK. This micro-AD plant was equipped with various components, including a manual shredder for biomass loading, a pre-digester tank (0.65 m^3) and a feed pump. Additionally, it featured a main anaerobic digester tank (2 m^3) with an automated mechanical mixer and a heater powered by an internal water heat exchanger. Among its other key components were a hydrogen sulphide scrubber filled with activated carbon pellets, a floating gasometer for biogas storage, a digestate sedimentation tank, and a digestate liquor storage tank.

The preliminary assessment, see supplementary materials, reveals that the feed composition consists of various components, including apples, coffee waste, green waste, catering waste, waste oats, soaked

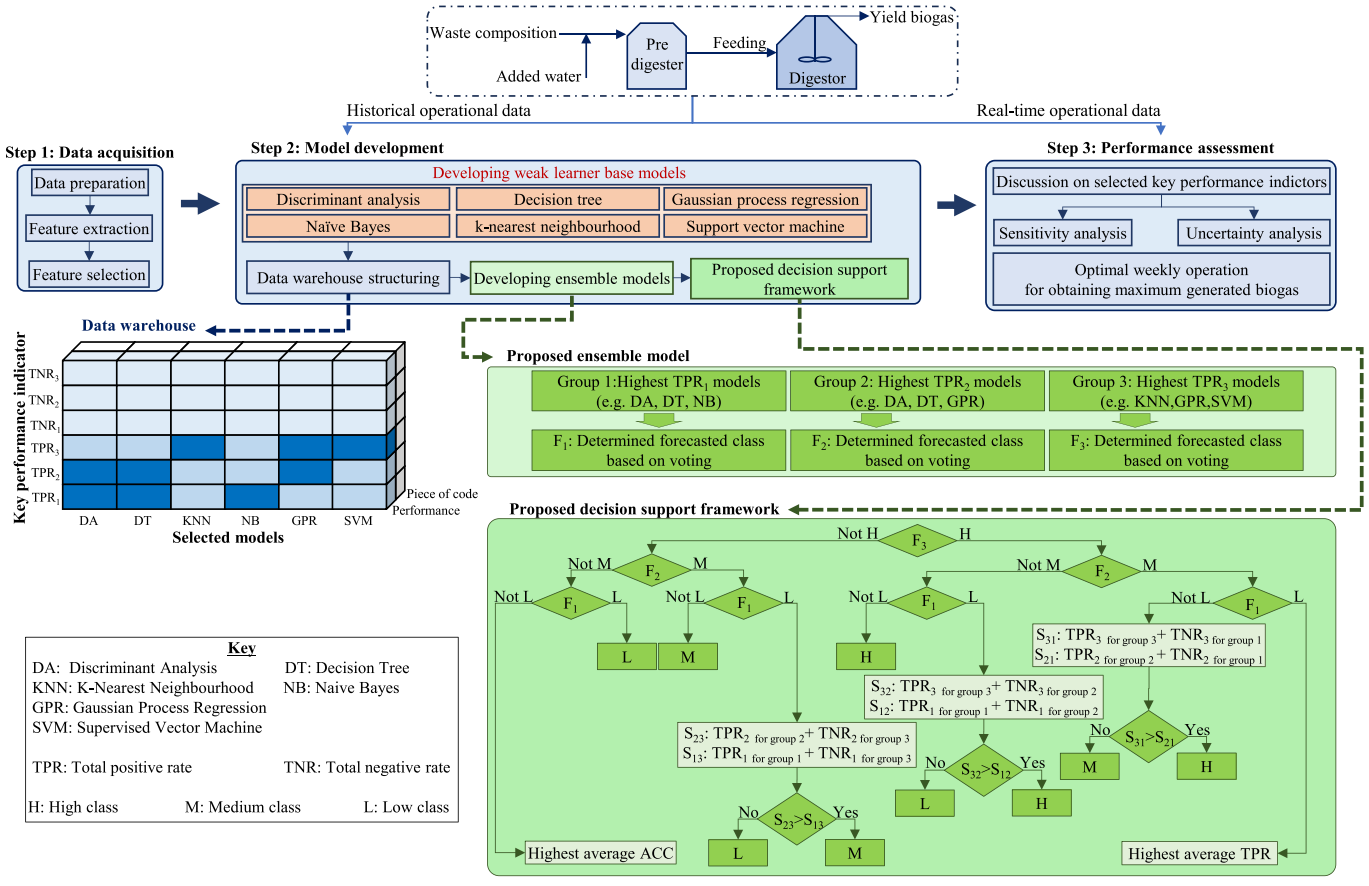


Fig. 1. The proposed framework for developing an ensemble model for real-time operation of anaerobic operation to maximise biogas production.

peanuts, tea leaves, tea bags, oil, soaked muesli and soaked caddy liners. However, it was observed that only catering waste, oats, and soaked liners constitute the major portion of the received daily waste. Furthermore, the analysis indicates a stable situation in terms of total solids, volatile solids and digester temperature, rendering them potentially removable from the decision variables.

Hence, the feed composition and AD output are grouped into six main features (data sets) as feed (FF), catering (CF), oat (OF), liner (LF), water (WF) and biogas (BF). Each data set is transformed into four distinct classifications: negligible (N), low (L), medium (M), and high (H) classes as listed in Table 1. The range of each class is determined based on operational practices and desired industrial goals using the *k*-mean classification model. For each group of data, seven features are introduced, each representing one specific day of operation, which characterise the 7-day operation of the AD system. For instance, concerning the feeding data (FF), features FF_t to FF_{t-6} are constructed, where FF_t represents the feeding volume of the current day, and FF_{t-6} represents the feeding volume of six days ago.

The extracted features were subjected to refinement using two established techniques including partial least squares (PLS) and sequential sensitivity (SA) analyses. More comprehensive explanations of these techniques is provided in the results section. These techniques are widely recognised as essential preliminary steps for identifying key variables that enhance classification performance while reducing computation times (Khan et al., 2023). PLS is used to estimate linear relationships between dependent and independent variables, revealing the direct effect of independent variables on the dependent variables (Orzi et al., 2018). The selected features, after undergoing these refinement techniques, were stored in the data warehouse for the subsequent development and testing using the WLDM models.

Table 1
Group features of feed composition and AD output extracted for developing weak learning data mining models with the classification rates identified for the micro-AD of this study.

Group feature*	Data unit	Description	Defined Classification
Feed (FF)	kg/day	The fresh weight of organic material sent to the digester for decomposition, measured in kg.	N: Zero L: <20 M: 20–40 H: >40
Catering (FC)	% Daily pre-feeding**	The composition of food waste produced in commercial kitchens, canteens, and restaurants.	N: Zero L: <16 % M: 16–50 % H: >50 %
Oat (FO)	% Daily pre-feeding	The composition of organic wastes containing oat grains	N: Zero L: <9 % M: 9–45 % H: >45 %
Liner (FL)	% Daily pre-feeding	The composition of organic wastes containing liners soaked in liquid.	N: Zero L: <7 % M: 7–63 % H: >63 %
Water (FW)	% Daily pre-feeding	The amount of water added to the pre-digester during its operations.	N: Zero L: <5 % M: 16–40 % H: >40 %
Biogas (FB)	m^3/day	The end product of the entire anaerobic digestion process measured in volumes.	L: <1 M: 1–4 H: >4

2.2. Step 2: Model development

The development of WLDM models comprises six different techniques: discriminant analysis (DA), decision tree (DT), *k*-nearest neighbourhood (KNN), naïve Bayes (NB), Gaussian process regression (GPR) and support vector machine (SVM). These specific models were selected based on their widespread application and recognised potential in previous anaerobic digestion processes, where they have been applied for various purposes (Cruz et al., 2022; Gupta et al., 2023; Khan et al., 2023). Each model was developed using MATLAB 2022b and optimised through automatic hyperparameter optimisation, including linear, Gaussian, kernel, quadratic, or course forms, that aims to minimise the five-fold cross-validation loss over 30 iterations. The five-fold cross-validation method was adopted to mitigate the error bias (Offie et al., 2023). For further details on the optimisation process, see the supplementary materials.

The dataset was divided into three distinct portions for training and testing the models based on recommendation of Piadeh *et al.* (2023). More specifically, 60 % of the dataset was allocated for training the individual WLDM models. Another 20 % of the dataset was reserved for testing the performance of these models. The remaining 20 % of the dataset was set aside as unseen data for evaluating the proposed ensemble model. To ensure equal representation of the databases, the group features were randomly distributed across the training, validation and testing databases. This approach aimed to maintain a balanced and representative distribution of data. Subsequently, the built WLDM models were stored in a model library and their KPIs were stored in the data warehouse cube shown in Fig. 1.

The KPIs of the models developed to predict biogas production were assessed using the confusion matrix concept as a statistical classification technique (Grandini et al., 2020; Tharwat, 2021). This technique involves mapping the three predicted biogas classes (i.e., low, medium, high) onto the confusion matrix. Using this mapping of the confusion matrix, this study employed two main KPIs of true positive rate (TPR) i.e., the ratio of correct prediction of the i^{th} class of yielded biogas and true negative rate (TNR) i.e., the ratio of correct rejection for a situation in which the biogas yield is not within the i^{th} class. These two KPIs are determined for each of the classes of yielded biogas as class 1 (low), class 2 (medium) and class 3 (high). As such, TPR and TNR are determined based on Eqs. (1) and (2). Model library and data cube are integrated as a data warehouse used for developing the ensemble model.

$$TPR_i\,(\%) = \frac{TP_i}{n_i} \times 100 \tag{1}$$

$$TNR_i\,(\%) = \frac{TN_i}{n_i} \times 100 \tag{2}$$

where TPR_i is the TPR of the i^{th} class; TNR_i is the TNR of the i^{th} class; TP_i is the number of the correct i^{th} class prediction; TN_i is the number of correct rejections of non– i^{th} class prediction and n_i is the total number of measured i^{th} classes.

The proposed ensemble model was developed by combining the developed WLDMs to create a more robust and accurate prediction model. The stacking method was selected due to the homogeneous nature and high variance and bias of the data (Kazemi *et al.,* 2020). This method involved the training of all WLDM models on the same set of training data. The WLDM models were blended afterwards using the proposed framework, hereafter denoted the 'smart model', which uses a decision support framework inspired by the bucket of models.

Sets of WLDM models are adjusted first to provide three classes as shown in Fig. 1 (See three groups mentioned in the proposed ensemble model in Fig. 1). Each group consists of higher performance in each of the KPIs that are previously stored in the data cube. For example, group 1 models are the models in which TPR_1 is recorded in the range of acceptable (e.g., DA, DT, and NB model in Fig. 1). To screen and exclude low-performance models, here, the performance rate is selected as 70 %

based on the recommendations of Cruz et al. (2022), and Khan et al. (2023). After determining the forecasted biogas class by all selected WLDM in each group, the answers are blended by hard voting techniques in which the most predicted class becomes the ensemble's prediction. As a result, a total of three answers i.e., F_1, F_2 and F_3 in Fig. 1, are generated by this ensemble modelling.

The predicted answers are then fed into a decision support framework as shown in Fig. 1 to determine the final prediction. This framework operates under specific conditions to identify the most appropriate predicted class. In scenarios where a single predicted class aligns with the selected group, that particular class is chosen as the final prediction. To illustrate this process, consider an example where group 3 predicts a high value for F_3 ($F_3 = H$), and F_2 is predicted to be anything other than medium ($F_2 =$ L or H, but not M), and similarly, F_1 is predicted to be anything other than low ($F_1 =$ M or H, but not L). Under these conditions, the correct predicted class would be H, following the fifth left branch of the flowchart in Fig. 1. On the other hand, if all models predict their respective classes, the model with the highest average of TPR takes precedence. This is demonstrated in the first right branch of the flowchart in Fig. 1. This approach ensures that the model with the highest correct accuracy is selected when all models agree on their predictions.

In cases where none of the models can accurately predict their respective classes, the final decision is made by assessing the overall performance of these models using the highest average of TPR and TNR. This criterion is represented by the first left branch of the flowchart in Fig. 1. When faced with situations where two models strongly advocate for their respective classes and are unable to reach a consensus, the model with the highest Score value (S_{ij}), as determined based on Eq. (3), is selected as the final decision.

$$S_{ij} = \mathrm{TPR}_{\,iforgroupi} + \mathrm{TNR}_{\mathrm{i\ for\ group\ j}} \tag{3}$$

where S_{ij} is the determined score, i and j are the two selected groups.

For this purpose, two scores are determined. For example, F_3 and F_2 are predicted as H and M, respectively, and F_1 is predicted as not L. In this scenario, the first score is calculated as the summation of the TPR of group 3 and the TNR of group 1 (S_{31} in the second right branch of the flowchart in Fig. 1). The other score is computed as the summation of the TPR of group 2 and the TNR of group 1 (S_{21}). To make the final decision, the two scores are compared. If S_{31} is greater than S_{21}, F_3 is selected as the final prediction. Otherwise, F_2 is chosen. This approach effectively evaluates the capability of true prediction of the two-group model (consisting of groups 1 and 2 in this example) based on the TNR rate of the other model group (group 3). By employing this decision support framework, the model systematically determines the final prediction in cases involving multiple predicted classes from different groups. This method ensures a structured and reliable approach to arrive at the most suitable prediction based on the aligned groups' predictions.

2.3. Real-time weekly operation

The model was optimised by using shuffled frog leaping algorithm (SFLA) to determine the optimum condition for the weekly operation of the AD plant for maximum biogas generation. The model was optimised for a 7-day ahead cycle to provide the operator with the optimal weekly operational pattern of the AD plant. To achieve this objective, a set of constraint rules can also be incorporated into the process of selecting the optimal scenario. These constraints serve as guiding principles to determine the best course of action, considering various factors.

The following key constraint rules have been established and can serve as generalised recommendations for other similar projects, seeking to apply these principles to optimise patterns in their research or practical projects: (1) minimising input loads to ensure that the highest possible biogas yield is obtained for each unit of added material, optimising resource utilisation. (2) minimising collection days to mitigate the operational costs associated with material handling, transportation and processing, (3) minimising added water load aligning with the goal of conserving water resources and mitigating associated energy costs, resulting in more sustainable and efficient operation, (4) minimising feeding days for cost savings arise from frequency of operational activities and associated resource consumption.

By integrating these constraint rules, the optimisation framework aims to obtain a balance between maximising biogas production, minimising resource inputs and optimising operational costs. This comprehensive approach ensures that the chosen scenarios align with both environmental sustainability and economic efficiency goals. By understanding and following this pattern, operators can effectively optimise their operations to maximise biogas yield. The proposed pattern is evaluated against the conventional operation of micro-AD for a one-month period.

2.4. Step 3: Performance assessment

To assess the performance of ensemble models, in addition to TPR and TNR, accuracy or correct prediction of all classes (ACC), false positive ratio (FPR) i.e., the portion of abnormal prediction, overestimation rate and underestimation rate were calculated as shown in Eqs. (3) to (6). Moreover, the SA analysis was carried out to determine the impact of each feature on biogas generation. To do this, one feature at a time was removed, and the accuracy difference of the developed WLDM models was measured. Furthermore, this method served the purpose of providing insights into the impact of each feature on the proposed ensemble model. Uncertainty analysis was also carried out to show changes in the relative accuracy with corresponding reductions in the dataset (Piadeh et al., 2023).

$$\mathrm{ACC}\,(\%) = \frac{\sum \mathrm{TP_i} + \mathrm{TN_i}}{\sum \mathrm{n_i}} \times 100 \tag{3}$$

$$\mathrm{FPR_i} = \frac{\mathrm{FP_i}}{\mathrm{TN_i} + \mathrm{FP_i}} \tag{4}$$

$$\mathrm{Overestimation}\,(\%) = \frac{\sum \mathrm{FP_i}}{\sum \mathrm{n_i}} \times 100 \tag{5}$$

$$\mathrm{Underestimation}\,(\%) = \frac{\sum \mathrm{FN_i}}{\sum \mathrm{n_i}} \times 100 \tag{6}$$

where $\mathrm{FPR_i}$ is the FPR of the i^{th} class, $\mathrm{FP_i}$ is the portion of the situation in which i^{th} class is predicted as higher yielded biogas, $\mathrm{FN_i}$ is the portion of the situation in which i^{th} class is predicted as lower yielded biogas.

3. Results and discussion

Several benchmark models inspired by Gupta et al. (2023) were developed to facilitate a comparative analysis and served as a valuable reference point for evaluating the effectiveness of the models. These models were trained, validated and tested with the same dataset that was used for the developing the proposed framework. The benchmark models used in this study include: (1) a "hard voting" stacked model that specifies the final class based on the majority class label predicted by the individual WLDM models, and (2) a "soft voting" stacked model that specifies the probabilities or confidence scores assigned by each WLDM model for each class and blending them to predict the final class. It should be mentioned that low-performance models are excluded in both benchmark models based on the aforementioned criteria in the section 2. Furthermore, optimised stacking models were created by combining the best performance developed WLDM models (screening and excluding the low-performance model), including (1) ensemble of the best performance WLDM model in $\mathrm{TPR_1}$ i.e., low based, (2) ensemble of the best performance WLDM model in $\mathrm{TPR_2}$ i.e., medium based, and (3) ensemble of the best performance WLDM model in $\mathrm{TPR_3}$ i.e., high based.

Moreover, to ensure the generalisation and comprehensiveness of the

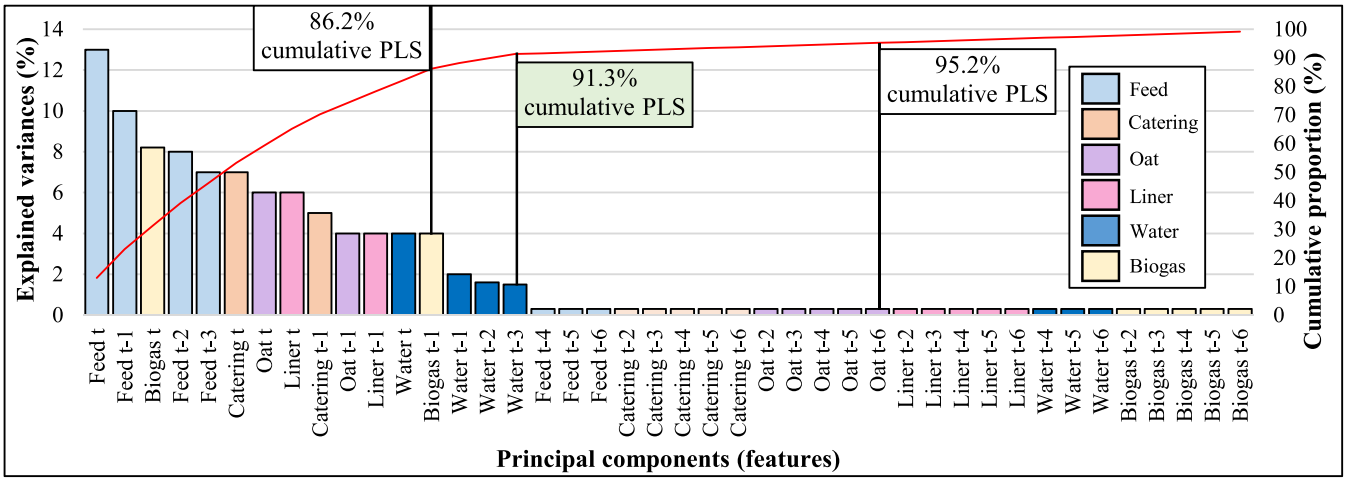
proposed model, other blending methods such as bootstrap aggregating (bagging) and boosting were also selected, including optimised versions of (1) RF, (2) subspace of developed NB, (3) XGBoost, (4) Gentle boost of developed DA model and (5) random under sampling and boosting (RUS Boost) of the GPR model. The SFLA optimisation technique, along with the classification and optimisation toolboxes of MATLAB 2022a, were employed to identify the best type of documented various models of stacking, bagging and boosting. During the optimisation, the number of learners was varied from 1 to 500, the learning rate ranged from 0.001 to 1, the maximum number of splits varied from 1 to 18,618 and the classification error improvement threshold was set at 0.01 % (Offie et al., 2023).

3.1. Feature analysis

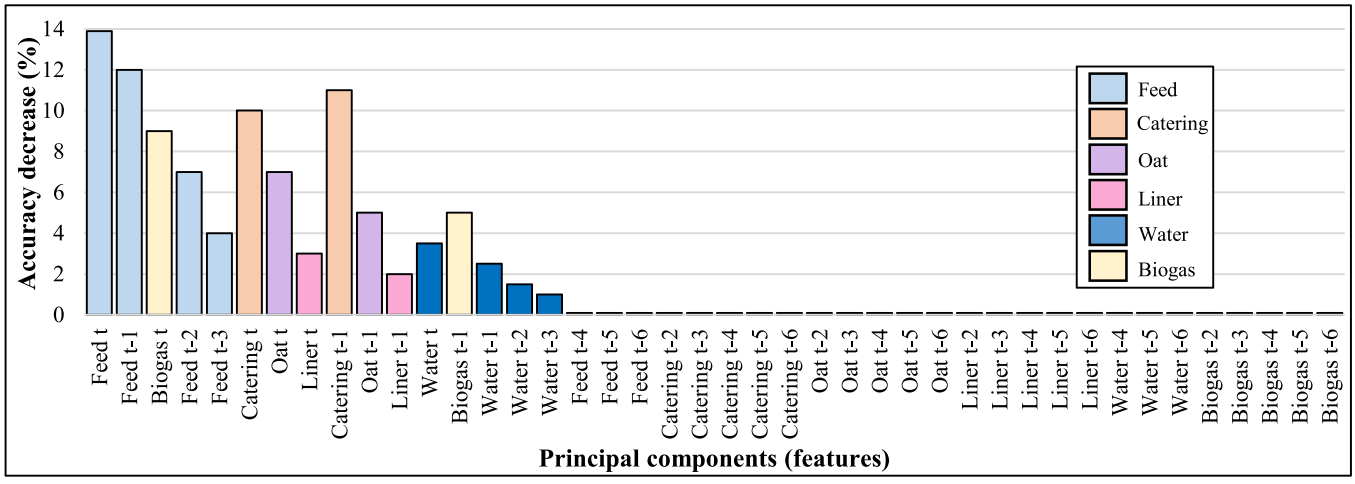
Fig. 2 presents the results of feature analysis for all the features outlined in Table 1. Out of the 42 total time-series features (i.e., 6 materials including FF, CF, OF, LF, WF and BF over the last 7 days), 16 features account for over 90 % of the cumulative variances in the PLS analysis (91.3 % in Fig. 2a). Notably, the feeding (FF) of the last four days (between day t and day/t^{-3}(−|-)) contributed significantly to this group of features, indicating its high impact on the modelling process. This observation is further corroborated by the SA analysis shown in Fig. 2b. The biogas levels (BF) at day t and t-1 also exhibit a substantial impact on the modelling outcomes i.e. biogas production at date t + 1, suggesting that some part of biogas production may be influenced by the feeding activities of the current day and the day before for the prediction of the next day's biogas yield. Moreover, the analysis demonstrates that the waste composition i.e. CF, OF and LF over the last two days significantly affected both the PLS analysis and the accuracy of WLDM modelling.

Also note that the overall importance of the features identified in the PLS analysis appears to align well with the results of the SA analysis, specifically when considering the criterion of cumulative PLS over 90 %. This implies that the features selected based on their cumulative PLS values above 90 % indeed have a significant impact on the accuracy of the model. However, it is essential to recognise that changing the criterion for cumulative PLS may lead to different results. For instance, if the criterion is set to cumulative PLS over 95 %, additional features with negligible impacts on accuracy may be included in the analysis, making the model less efficient in practice (as can be seen from a comparison of Fig. 2a with Fig. 2b). On the other hand, if the cumulative PLS criterion is lowered to 85 %, relevant features with a noticeable impact on accuracy, such as the three features related to added water, might be excluded, leading to potential loss of predictive power.

Therefore, relying solely on the PLS analysis, which is commonly utilised in several research works, might not always yield the most accurate or appropriate results. It becomes evident that incorporating the



(a)



(b)

Fig. 2. Feature analysis of the feed composition and AD output: (a) PLS analysis of extracted features, (b) average accuracy decrease of all WLDM models obtained by the SA analysis.

SA analysis mentioned earlier, is crucial in the feature selection process. Such an approach allows for a more robust understanding of the features' actual influence on the model's accuracy and helps in making informed decisions regarding their inclusion or exclusion of any feature for further analysis. By adopting this approach, the final model can be more reliable and better suited for practical applications.

Consequently, among all extracted time-series features, the appropriate class of biogas yield (low, medium and high) for the next day is predicted based on the following 16 features: (1–4) feeds to the main digester tank over the last four days (FF from day/t^{-3}(−|-) to day *t*); (5–8) water added to the pre-digester tank over the last four days (WF from day/t^{-3}(−|-) to day *t*); (9–10) biogas generation over the last two days (BF from day/t and day *t*), and (11–16) oat, liner and catering added to the pre-digester tank over the last two days (OF, CF and LF from day/t to day *t*). This structure of input and output is used for the development of individual WLDM models and the proposed ensemble models that will be presented and discussed in the following section.

3.2. Performance of individual WLDM models

The performance of six individual WLDM models is depicted in Fig. 3, based on three KPIs (TPR, TNR and ACC). For the TPR of the low class (Fig. 3a), the DA, DT and NB models demonstrate an acceptable rate, indicating their ability to correctly identify instances in this class. In the case of the medium class (Fig. 3b), the GPR model outperforms the NB model. Furthermore, for the high class (Fig. 3c), most models, except GPR, exhibit excellent performance in recognising situations with high biogas yields.

On the other hand, when comparing Fig. 3d – f, it becomes evident that although the models excel in detecting the high-class situation, they also display many instances of underestimation and overestimation in other situations, as evident from the low TNR in the low class and high class, which are only relatively compensated for in the medium class (Fig. 2e). These limitations result in low overall accuracy for all models, even in the best- performance one, DA, as shown in Fig. 3g, where it achieves an accuracy of only around 80 %. Overall, these findings

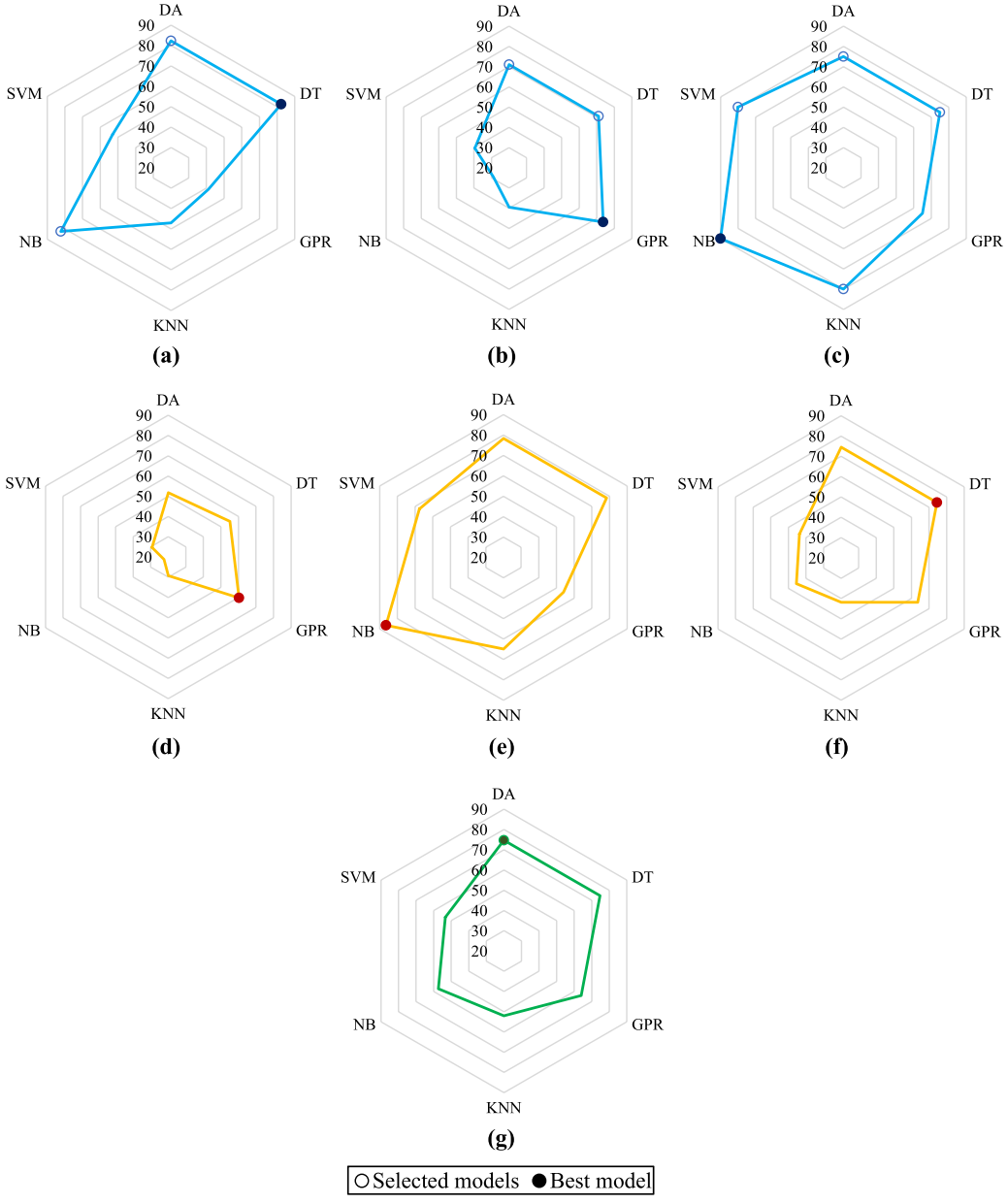


Fig. 3. Performance assessment of the individual WLDM models based on: (a) TPR of low class, (b) TPR of medium class, (c) TPR of high class, (d) TNR of low class, (e) TNR of medium class, (f) TNR of high class, (g) ACC rate.

suggest that while the selected models demonstrate proficiency in certain areas, they still suffer from some shortcomings that hinder their overall accuracy. Consequently, the proposed ensemble model combined the strengths of the superior models in each class and improved performance in a broader range of scenarios.

3.3. Performance of the ensemble models

Fig. 4 and Table 2 show the performance of the ensemble models, including the proposed ensemble and other benchmark models (further details given in Fig. A3). Compared to the performance of individual

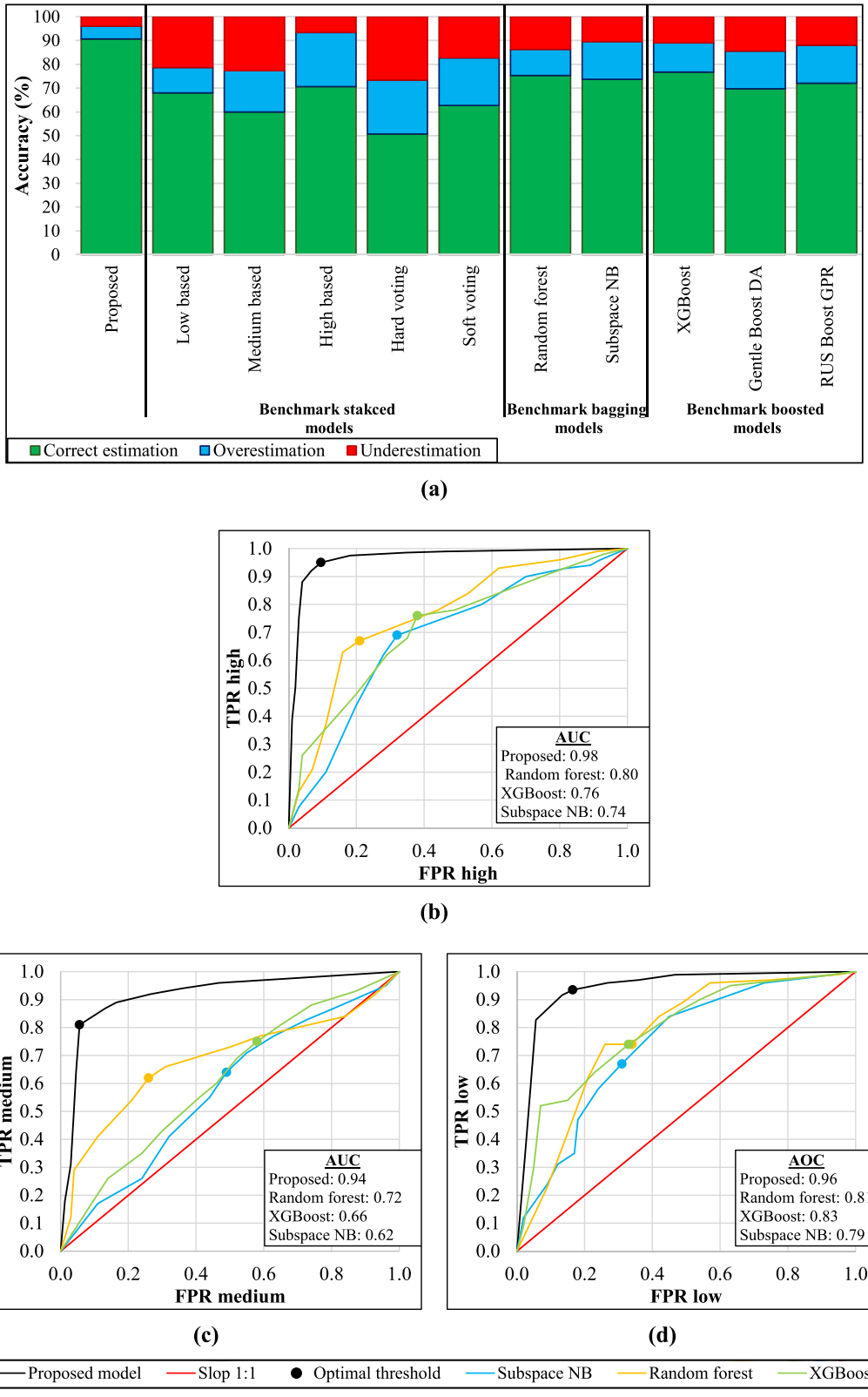


Fig. 4. Performance assessment of FPR for the proposed ensemble model compared with the three best performing benchmark ensemble models: (a) accuracy assessment, (b-d) ROC curves and AUC of the four best performance models, (b) high class, (c) medium class, (d) low class.

WLDM models, the accuracy of the ensemble models improved from 80 % to 91 % in comparison to the best performance WLDM model (DA model in Fig. 3g). Fig. 4a also shows that the proposed ensemble model outperforms other benchmark models with a remarkable 5 % for each underestimation and over estimation. As can be seen, both hard and soft voting approaches of all stacked models showed an accuracy around 63 % and 51 % respectively which indicates the relatively low accuracy of all individual WLDM models. On the other hand, group stacking of the models based on their capability in a specific class – for example, the low or high class – shows a better result, especially for a high-based model, which could improve accuracy to near 70 %. However, large overestimation of this model still is challenging especially for optimal operation in which higher rate of yielded biogas is a goal.

The results of comparing ensemble models reveal the high accuracy of bagging and boosting models, particularly Random Forest and XGBoost, over benchmark stacking models. This finding aligns with previous research conducted on numerical problems (Xu et al., 2021; Sonwai et al., 2023). However, it is important to note that despite this progress, the obtained accuracy, which remains below 80 %, still falls short when compared to the performance of the proposed stacked model. To further validate these outcomes, the study examines the receiver operating characteristic (ROC) curves and their corresponding area under the ROC curve (AUC) for the top four best performance models. The proposed ensemble model notably excels by consistently maintaining an AUC above 0.94 across all three classes, particularly demonstrating exceptional performance in the high class where the AUC is reported as 0.98 (Fig. 4b). In contrast, the alternate benchmark models exhibit AUC values ranging from 0.74 to 0.8 in the high class, while their performance notably deteriorates in the medium class with AUC figures of 0.62 to 0.72. Moreover, it is noteworthy that the optimal thresholds for the proposed model remained relatively consistent along the x-axis, within the range of 0.8–1 in TPR rates and 0.2–1 for FPR rates, while for the other models, these thresholds shifted towards higher values. These observations provide strong evidence of the better performance of the proposed model than the other benchmark models.

3.4. Further analysis

Fig. 5 illustrates the results of the uncertainty and sensitivity analysis conducted on the developed model for biogas prediction. As depicted in Fig. 5a, the models exhibit a notable ability to be trained using a reduced dataset size of up to 80 % of the total available training data while maintaining a remarkable 95 % accuracy. In simpler terms, the models showcase robust performance within this specified range of training dataset size, displaying resilience against the impact of dataset reduction. However, it is important to note that as the dataset size is further reduced beyond this range, the models exhibit an adaptive behaviour resulting in a gradual nonlinear decline in accuracy. Notably, as the dataset size dwindles to less than 30 %, the models encounter significant challenges, ultimately leading to a complete failure in performance, with accuracy levels plummeting to nearly 0 %. This particular revelation bears substantial significance. Despite the models being developed within a relatively uncomplicated framework utilising input data spanning close to a year, the demonstrated adaptability and efficiency within this context can have far-reaching implications. Such efficiencies have the potential to yield substantial energy cost savings and mitigate the need for recurrent and time-consuming retraining, particularly in the context of broader industrial applications.

The sensitivity analysis focusing on the influence of removing individual groups of features is presented in Fig. 5b. The observed decline in overall accuracy underscores the pivotal role of specific group features, particularly the feeding-related attributes, in shaping the model's performance. Notably, the removal of these feeding-related features results in a substantial 50 % drop in accuracy, contributing significantly to both overestimation and underestimation tendencies. Interestingly, the nature of impact varies across different group features. Specifically, the removal of catering and oat-related group features primarily leads to an increase in underestimation, while the attributes related to biogas, water and liner exert a more pronounced effect on overestimation tendencies.

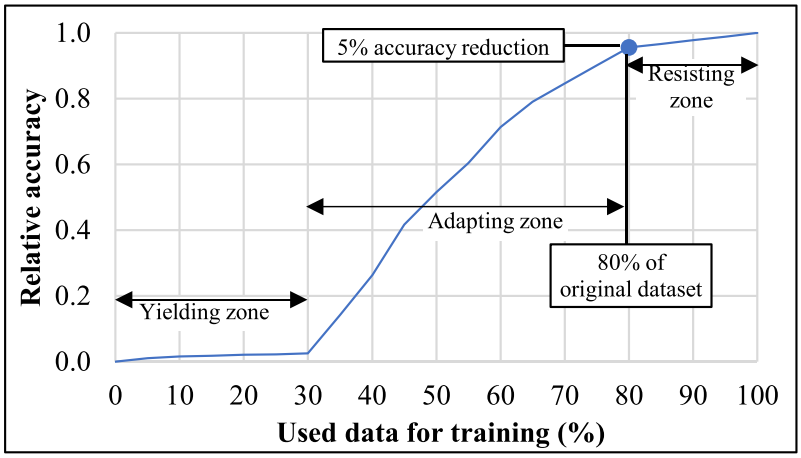
This outcome underscores the unique role of each material exerting a distinct influence on the model's predictive performance. Moreover, insights derived from previous biogas production provide valuable cues to the model regarding the residual potential for biogas release. This, in turn, has the potential to mitigate overestimation in future predictions. Remarkably, the incorporation of this group of features as input has the capacity to alleviate a significant portion of overestimation instances, effectively addressing approximately 20 % of such cases.

3.5. Optimal operational pattern for maximum biogas generation

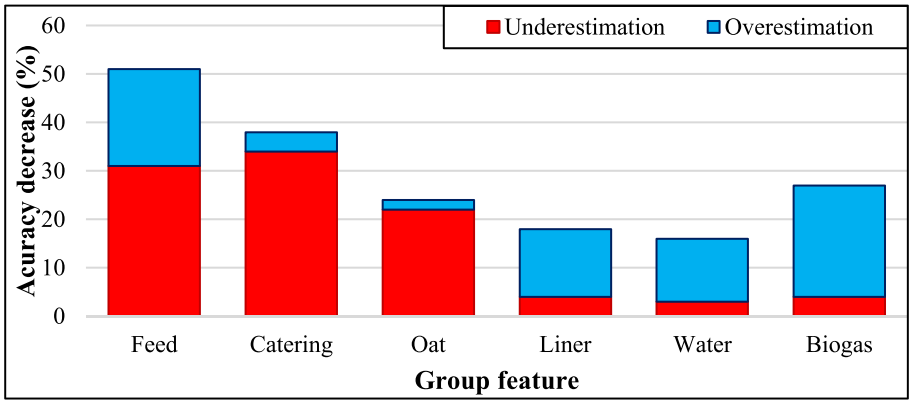
Fig. 6a presents the optimal weekly condition and best input pattern for obtaining maximum volume of biogas from the AD plan for the case study. This pattern undergoes rigorous testing with previously unseen data spanning a period of 76 days. A comparative analysis is then conducted against the actual operational performance recorded most productive phase of biogas generation during this period. Fig. 6b clearly demonstrates the efficacy of the proposed optimal pattern in increasing the number of days with high biogas yield. According to the data, the

Table 2
Performance of the developed ensemble models.

Model	ACC	TPR class			TNR class		
		Low	Medium	High	Low	Medium	High
Proposed	91	90	88	93	91	92	89
Benchmark staked models							
Low based	68	85	72	53	62	66	78
Medium based	60	50	84	47	64	48	69
High based	71	50	68	87	78	72	60
Hard voting	51	30	76	43	58	38	56
Soft voting	63	45	56	80	69	66	51
Benchmark bagging models							
RF	75	83	72	73	73	77	77
Subspace NB	74	65	65	87	77	78	65
Benchmark boosted models							
XGBoost	77	78	76	77	76	77	77
Gentle Boost DA	70	78	48	82	67	80	60
RUS Boost GPR	72	55	72	83	78	72	64



(a)



(b)

Fig. 5. Further analysis on proposed ensemble model: (a) uncertainty analysis on size of dataset, and (b) sensitivity analysis on group features.

implementation of the proposed optimal pattern can lead to a notable 78 % increase in the duration of time which substantial biogas production is achieved. This significant improvement underlines the effectiveness of the proposed approach in optimising biogas generation.

From Fig. 6a, it is observed that the feed feature is heavily fed into the digester only on the 4th and 7th days. Catering feature needs to be fed into the pre-feed tank on the 3rd, 4th, 6th and 7th days. The oat feature is fed into the pre-feed tank on the 3rd and 6th days. Water is added to the digester on the 3rd, 4th, 6th and 7th days to obtain maximum biogas. Intriguingly, the absence of the liner input in the optimal condition signifies its negligible positive impact on enhancing biogas generation. This finding underscores that the inclusion of the liner feature has no significant contribution to the overall biogas yield, thus making its omission from the optimal setup a justifiable decision.

In the case of the feed feature, high-class variables are strategically incorporated into the digester for a mere two days within the week. This observation underscores a potential strategy for optimising biogas generation, indicating that rather than continual input of materials, a more effective approach could involve extending the intervals between feed additions by 2 or 3 days, followed by a substantial surge in the system's load. This noticeable difference becomes apparent in Fig. 6c, where a clear departure from the usual practice of frequent waste feeding into the pre-digester (black dots) is vividly reduced by implementation of the proposed strategy (red dots). Results highlight a substantial and statistically significant reduction of 71 % in the amount of time spent on operational activities, during which the mechanised pumping mechanism facilitates the controlled transfer of materials into the digester. This reduction carries important implications, encompassing energy conservation, as well as a notable decrease in the demands for careful monitoring, extensive maintenance efforts and labour costs.

Interestingly, the manner in which catering and oat materials are suggested showcases contrasting patterns. The model proposes an initial infusion of a substantial quantity of catering materials into the pre-digester, followed by a subsequent day with a relatively low catering input. Conversely, a light input of oat material on one day is suggested. Furthermore, the recommended approach for adding water demonstrates a distinct trend. It also suggests that the addition of water should be prioritised on days when waste materials are input (as evident on days 3, 4, 6 and 7 in Fig. 6a) that is compatible with experimental experiences in which pumped water used for dilution of dry feedstock or for cleaning the hammer mill before the pre-feed tank.

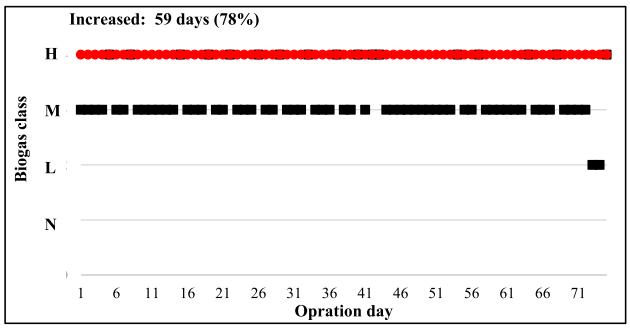
Upon comparing the input pattern with the real case observations, see the supplementary materials, it becomes evident that the total operation days for each input increased by approximately 25 %. However, when considering the overall picture, as depicted in Fig. 6d, there is a 30 % decrease in the total number of pre-feeding days. This indicates that despite the individual increase in the number of input materials being added, the strategy of compacting them on specific days contributes to a reduction in the overall operation days and associated costs.

This pattern serves to illustrate the applicability and validity of the developed prediction model when coupled with the proposed optimisation method. While it is understandable that the demonstrated pattern may not be applicable directly in other research or industrial AD projects, the underlying concept can be employed on other projects to achieve similar conservation, especially on operational days and with

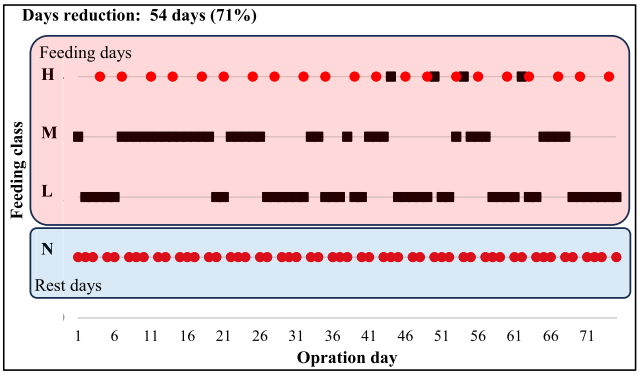
Features	Timeline (day)						
	t+1	t+2	t+3	t+4	t+5	t+6	t+7
Feed	N	N	N	H	N	N	H
Catering	N	N	H	L	N	H	L
Oat	N	N	L	N	N	L	N
Liner	N	N	N	N	N	N	N
Water	N	N	L	L	N	L	L

N: Negligible L: Low M: Medium H: High

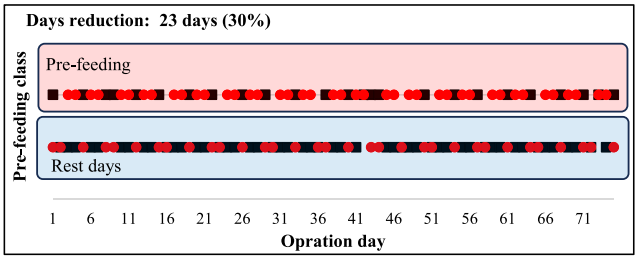
(a)



(b)



(c)



(d)

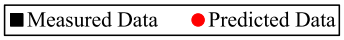


Fig. 6. Comparison between optimal weekly operation pattern and testing event (76 days): (a) suggested optimum condition for the operation of the micro-AD plant for maximum biogas generation, (b) biogas yield, (c) feeding to digester, and (d) pre-feeding days.

other resources/materials.

4. Limitations and future perspective

This study holds significant potential in accurately predicting biogas and optimising real-time operations in AD projects. However, certain limitations need to be acknowledged, pointing towards avenues for further research and development. A primary limitation lies in the challenge of accessing comprehensive big data and operational databases for time spans shorter than a day and datasets spanning over a year. This challenge arises due to commercial-in-confidence of large AD projects or the relatively nascent nature of using micro-AD projects. While parameters such as the volume of added water, daily feeding and generated biogas can be automatically tracked with high reliability, the composition of waste materials continues to rely on operator reports, introducing uncertainties into the data.

To overcome these limitations and advance the proposed model, a series of pre-processing steps such as harnessing data mining techniques and bolstering the capabilities of real-time remote sensing can be considered. Moreover, it is imperative to subject both the proposed model and the distinct weekly pattern introduced in this study to longer-term testing across various timeframes and within comparable projects. Such extended validation efforts would provide a comprehensive understanding of the model's effectiveness and potential scalability. An intriguing avenue for exploration involves integrating the real-time operational pattern with risk assessment. These could encompass risk scenarios such as shifts in waste composition or errors made by operators while adding input materials. The introduced pattern and optimisation framework have the potential to dynamically adapt the weekly pattern to address these operational challenges, suggesting a broader application within the realm of digital visualisation projects.

Finally, this study predominantly relied on WLDM models, with the exclusion of more advanced deep learning or recurrent data-driven modelling techniques from its scope. Nevertheless, even though the application of these alternative models has been tested within the context of AD design, planning, and operation (as referenced in Gupta et al., 2023), it is imperative that the accuracy of the proposed model be compared and validated against the approaches outlined in such studies.

5. Conclusions

This study introduces an Ensemble-based framework that offers a suggested real-time weekly operation pattern aiming to enhance biogas generation in an AD plant. The proposed model exhibits a remarkable 91 % accuracy in providing accurate estimations, outperforming other developed models achieving 50–80 % accuracy. Both PLS and SA analyses reveal a high sensitivity to the feed feature. The optimised weekly AD operation demonstrates promising results, with a substantial 78 % increase in the number of days achieving high biogas generation, accompanied by a 71 % reduction in total required feeding days and a 30 % reduction in pre-feeding days.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors wish to thank Diego Vega and Rokiah Yaman from LEAP Micro AD and Dr Davide Poggio from the University of Sheffield for their great support to provide and analyse the data collected from the case study. The authors wish to thank the editor and the anonymous reviewers for making constructive comments which substantially improved the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2023.130017.

References

Almomani, F., 2020. Prediction of biogas production from chemically treated co-digested agricultural waste using artificial neural network. Fuel 280, 118573.

Awasthi, M., Yan, B., Sar, T., Gómez-García, R., Ren, L., Sharma, P., Binod, P., Sindhu, R., Kumar, V., Kumar, D., Mohamed, B., Zhang, Z., Taherzadeh, M., 2022. Organic waste recycling for carbon smart circular bioeconomy and sustainable development: A review. Bioresource Technology 360, 127620.

Clercq, D., Wen, Z., Fei, F., Caicedo, L., Yuan, K., Shang, R., 2020. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. Science of the Total Environment 712, 134574.

Cruz, I., Andrade, L., Bharagava, R., Nadda, A., Bilal, M., Figueiredo, R., Ferreira, L., 2021. An overview of process monitoring for anaerobic digestion. Biosystem Engineering 207, 106–119.

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

Emebu, S., Pecha, J., Janacova, D., 2022. Review on anaerobic digestion models: model classification & elaboration of process phenomena. Renewable Sustainable Energy Revision 160, 112288.

Fajobi, M., Lasode, O., Adeleke, A., Ikubanni, P., Balogun, A., 2022. Effect of biomass co-digestion and application of artificial intelligence in biogas production: A review. Energy Sources, Part a: Recovery, *Utilization, and Environmental Effects* 44 (2), 5314–5339.

Grandini, M., Bagli, E., Visani G. (2020). "Metrics for Multi-Class Classification: An Overview. *Computer Science*, ArXivabs/2008.05756.

Gupta, R., Zhang, L., Hou, J., Zhang, Z., Liu, H., You, S., Ok, Y., Li, W., 2023. Review of explainable machine learning for anaerobic digestion. Bioresource Technology 369, 128468.

Jia, R., Song, Y., Piao, D., Kim, K., Lee, C., Park, J., 2022. Exploration of deep learning models for real-time monitoring of state and performance of anaerobic digestion with online sensors. Bioresource Technology 363, 127908.

Kazemi, P., Bengoa, C., Steyer, J.-P., Giralt, J., 2021. Data-driven techniques for fault detection in anaerobic digestion process. Process Safety and Environmental Protection. 146, 905–915.

Khan, M., Chuenchart, W., Surendra, K., Khanal, S., 2023. Applications of artificial intelligence in anaerobic co-digestion: Recent advances and prospects. Bioresource Technology 370, 128501.

Masalegooyan, Z., Piadeh, F., Behzadian, K., 2022. A comprehensive framework for risk probability assessment of landfill fire incidents using fuzzy fault tree analysis. Process Safety and Environmental Protection 163, 679–693.

Offie I., Piadeh F., Behzadian K., Campos L., Yaman R. (2022). Real-Time monitoring of decentralized Anaerobic Digestion using Artificial Intelligence-based framework. *International Conference on Resource Sustainability (icRS 2022)*, pp. 1-4.

Offie, I., Piadeh, F., Behzadian, K., Campos, L., Yaman, R., 2023. Development of an artificial intelligence-based framework for biogas generation from a micro anaerobic digestion plant. Journal of Waste Management 158, 66–75.

Orzi, V., Riva, C., Scaglia, B., D'Imporzano, G., Tambone, F., Adani, F., 2018. Anaerobic digestion coupled with digestate injection reduced odour emissions from soil during manure distribution. Science of the Total Environment 621, 168–176.

Sappl, J., Harders, M., Rauch, W., 2023. Machine learning for quantile regression of biogas production rates in anaerobic digesters. Science of the Total Environment 872, 161923.

Sonwai, A., Pholchan, P., Tippayawong, N., 2023. Machine learning approach for determining and optimizing influential factors of biogas production from lignocellulosic biomass. Bioresource Technology 383, 129235.

Tharwat, A., 2021. Classification assessment methods. Applied Computing and Informatics 17 (1), 168–192.

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

Wang, Z., Peng, X., Xia, A., Shah, A., Huang, Y., Zhu, X., Zhu, X., Liao, Q., 2022. The role of machine learning to boost the bioenergy and biofuels conversion. Bioresource Technology 343, 126099.

Wang, Z., Peng, X., Xia, A., Shah, A., Yan, H., Huang, Y., Zhu, X., Zhu, X., Liao, Q., 2023. Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass. Energy 263 (D), 125883.

Xu, W., Long, F., Zhao, H., Zhang, Y., Liang, D., Wang, L., Larson, K., Cao, L., Zhang, Y., Liu, H., 2021. Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms. Waste Management 121, 59–66.

Yan, P., Gai, M., Wang, Y., Gao, X., 2021. Review of Soft Sensors in Anaerobic Digestion Process. Processes 9, 1434.

Yildirim, O., Ozkaya, B., 2023. Prediction of biogas production of industrial scale anaerobic digestion plant by machine learning algorithms. Chemosphere 335, 138976.