Diagnosis of undesired scenarios in hydrogen production by photo-fermentation
Isaac Monroy and Germán Buitrón

ABSTRACT
This study presents the use of a machine learning method from the artificial intelligence area, such as the support vector machines, applied to the construction of data-based classification models for diagnosing undesired scenarios in the hydrogen production process by photo-fermentation, which was carried out by an immobilized photo-bacteria consortium. The diagnosis models were constructed with data obtained from simulations run with a mechanistic model of the process and assessed on both modelled and experimental batches. The results revealed a 100% diagnosis performance in those batches where light intensity was below and above an optimum operation range. Nevertheless, 55% diagnosis performance was obtained in modelled batches where pH was away from its optimum operation range, showing that diagnosis model predictions during the first observations of those batches were classified as normal operation and revealing diagnosis delay in pH oscillations. In general, results demonstrate the reliability of classification models to be used in future applications such as the on-line process monitoring to detect and diagnose undesired operating conditions and take corrective actions on time to maintain high hydrogen productivities.

Key words | diagnosis models, hydrogen production, mechanistic model, photo-fermentation process, supervised learning methods, undesired scenarios

INTRODUCTION
Hydrogen is a green fuel with the highest energy content per weight that can be produced using renewable materials such as biomass (Kapdan & Kargi 2006; Guo et al. 2010) and water (Bartels et al. 2010) and whose combustion only generates water without carbon-based emissions. These traits support the relevance to look for more sustainable processes to produce it in large scale.

Among the biological processes for producing biohydrogen, photo-fermentation can be found, which is carried out by a group of anoxygenic photosynthetic bacteria named purple-non-sulfur bacteria (PNSB), capable of converting volatile fatty acids (VFA), organic acids and sugars into hydrogen with the aid of light energy (Chen et al. 2011). In this context, PNSB present the advantage to be able to use light into a wide wavelength range, which is included in the solar light spectrum.

Concerning the photo-fermentation process, it is still common to find in literature the use of pure cultures for producing biohydrogen, which is impractical in terms of energy cost and organic matter degradation for treating wastewaters. In contrast, mixed cultures or bacterial consortia represent better alternatives in these terms even though only a few applications on the bio-hydrogen production have been reported (JianLong & Wei 2008; Assawamongkholsiri & Reungsang 2015; Guevara-López & Buitrón 2013).

Even less effort has been made in using immobilized consortia despite their known advantages such as high biomass concentrations, high resistance to inhibitors (Guevara-López & Buitrón 2013), and higher H2 productivities than with suspended cultures (Levin et al. 2004; Argun & Kargi 2011).

Modelling hydrogen production is one of the most critical requirements of the process for predicting the hydrogen yield and scaling up the process (Prakasham et al. 2011; Nasr et al. 2013). Mechanistic models, also called mathematical or analytical, are quantitative models normally described by differential equations of the main process variables such as substrate, biomass and product concentration, which represent their trajectories through the process time.

Although mechanistic models try to fit real bioprocesses, sometimes it turns out a tough duty due to the several metabolic pathways involved or activated under certain process conditions, which are maybe unknown in that biological system. In contrast, data-based models are alternative models constructed from the existing process history and plant or laboratory records that can be useful not only for predicting product yields and productivities, but also for detecting and diagnosing process faults through a supervised learning from the process data.

Fault detection and diagnosis are of primary concern in both chemical and biological processes. In this sense, there have been many ways to address these challenges, such as the application of multivariate statistical techniques that help to monitor processes on real-time (Venkatasubramanian et al. 2003a), and the use of artificial intelligence (AI) techniques and supervised learning methods (machine learning area) to create data models able to detect and diagnose different process scenarios, most of them considered as faults (Monroy et al. 2012a, 2012b).

Supervised learning approaches applied for fault detection and diagnosis use prior knowledge on process faults given by patterns in the measured process variables with the assumption that process data can be correctly labelled in classes or faults (Venkatasubramanian et al. 2003b; Qin 2012). Most of the supervised learning methods are then based on classification techniques that detect and classify normal or faulty process scenarios based on pattern recognition into the classes that were previously learned and modelled (Ardakani et al. 2017).

Despite their great potential, data-based classification models have been barely used for fault diagnosis in only a few biological processes (Monroy et al. 2012a). Until now, they have not been applied to diagnose undesired scenarios in the hydrogen production process by photo-fermentation. Therefore, this kind of data driven models could be capable of monitoring the process and diagnosing abnormal scenarios given by changes in some process variables such as the increase or decrease in pH, light intensity (LI), or even both. Also, the diagnosis of these undesired but common scenarios may allow taking corrective actions on time that avoid a decrease in the hydrogen production.

This work proposes the construction and application of data-based classification models to the diagnosis of undesired scenarios in the hydrogen production process by photo-fermentation using an immobilized consortium of PNSB and incandescent lamps as artificial light source in the reactor. Classification or diagnosis models are constructed using a machine learning method, the support vector machines (SVM). It is worthy to mention that SVM have been successfully applied to fault detection and diagnosis in both chemical and biochemical processes (Monroy et al. 2012a, 2012b).

Data from the photo-fermentation process were obtained through simulations from a mechanistic model of the process, reported by Monroy et al. 2018. Classification models proposed in this work were validated onto both simulated and experimental data from the process to assess and compare the diagnosis performance of the undesired scenarios by the constructed models.

MATERIALS AND METHODS

A PNSB consortium, mainly composed of *Rhodopseudomonas palustris*, was immobilized on luffa fibres and used as biological system in laboratory experimental photo-fermentations. Isolation and identification of the microbial consortium, culture medium preparation as well as biomass immobilization are described in previous research (Guevara-López & Buitrón 2015).

Experimental photo-fermentations were set up in 100 mL serological bottles with 75 mL of culture medium at 6.7 initial pH and 300 mg volatile solids (VS) per litre. Bio-hydrogen was produced in these laboratory batch cultures using incandescent lamps, irradiated at several light intensities, and quantified by the water displacement method. Analytical methods used during photo-fermentations are clearly described in previous investigations (Monroy et al. 2016, 2018).

A mechanistic model of hydrogen production by photo-fermentation was previously constructed and documented by Monroy et al. (2018) based on such experimental information and using the kinetic parameters found by Zhang et al. (2015), who reported two mathematical models for *R. palustris* cultures, the predominant species in the utilized consortium, and found similar kinetic patterns to Monroy et al. (2018). Also, their cultures were irradiated with incandescent light during fermentation.

In the research reported by Monroy et al. (2018) two types of models were presented, both aimed at basically modelling the batch photo-fermentation process to obtain predictions of the product (hydrogen production). The first one was a mathematical model (mechanistic), used in this current work to obtain a great amount of process data needed for constructing any data-based diagnosis model. The second one was a black-box model constructed with the well-known method artificial neural networks (ANN).
and using some experimental data to predict hydrogen volumes along the process time.

Mechanistic models give prediction of the process based on values of some process variables such as the hydrogen production. In fact, the used mechanistic model (Monroy et al. 2018) renders values of chemical oxygen demand (COD), biomass (X) and hydrogen (H2) concentration each sampling time (20 h) up to the COD depletion. Many simulations were run subject to different initial conditions of COD concentration, biomass concentration, pH and LI to obtain a great amount of process data.

CLASSIFICATION MODELS

Two hundred and fifty simulations were run, shifting the input variable values. The whole information was gathered into a data matrix, which was composed of many rows containing the measurements at each sampled time (20 h) for all the 250 batches and six columns corresponding to the next monitored variables: process time (t), COD, X and H2 concentration, LI and pH. This data matrix represented the training data set needed for supervised learning in which a classification method is able to learn from this set and construct diagnosis models that would be capable to be used with further data to detect and diagnose the trained faults.

Unlike mathematical (mechanistic) and kinetic models, the outputs of diagnosis models are positive or negative values that indicate whether the process is behaving under normal operating conditions or not. If not, these models are able to diagnose which undesired scenario is occurring because some of them have been taught the classification method through the given process data.

Before constructing the classification or diagnosis models, principal component analysis (PCA) method was applied to the training data set. PCA is a multivariate statistical method used for process monitoring that was employed to find those operating conditions of LI and pH, for which the Q statistic values were over a threshold calculated with a 95% confidence level.

Q statistic, also called the squared prediction errors index, is calculated from PCA to quantify the lack of fitting between each observation and the PCA model and detect new events not captured by the model. It represents the squared distance from each observation to the residuals space (Montgomery 2005). In addition, the Q statistic limit is calculated based on the equation reported by Lu & Upadhyaya (2005).

Every process measurement with its Q statistic value above the Q limit then belonged to some batch operated under undesired conditions of either pH or LI and therefore is considered as faulty. These process conditions out of a nominal operation range for both pH and LI were then grouped as classes of undesired scenarios. More information about the use of PCA method for process monitoring can be found in Monroy et al. (2012a, 2012b).

Next, training data were arranged into four classes of undesired scenarios according to the results given by the PCA application. After that, data-based classification models were constructed as a result of applying the supervised learning method, SVM, to both the training data set and a label matrix with the same number of rows as the number of process measurements and four columns according to the number of classes. This label matrix supports the classification method to learn from the historic data and recognize patterns in them to construct models for each class of undesired scenarios, which would be then be used to diagnose the learned classes in the process.

SVM are a classification method from machine learning area developed by Vapnik and co-workers (Duda & Hart 1973), which is based on margin maximization between the training patterns from the training data and the decision boundary (classification models). Each classifier is based on two principal components: a weight vector (w) considered as a function, and the bias (b) which is the distance from the hyperplane to the origin. These two terms can be obtained from the next convex quadratic problem:

\[
\text{Minimize } ||w|| s.t. y_i((w, x_i) + b) \geq 1
\]

Therefore, the resulting classification models after applying SVM were tested on validation data sets obtained either from new simulations with the mechanistic model or from experimental photo-fermentations (Monroy et al. 2018) to assess the diagnosis performance of the four different classes of undesired scenarios in the hydrogen production process.

RESULTS AND DISCUSSION

Two hundred and fifty simulations under different conditions of pH and LI were run with the mechanistic model previously reported (Monroy et al. 2018), and used as training data set. PCA was applied to this data set, and four
principal components were extracted in the PCA model, which retained 98% cumulative variance from data.

The training data were projected onto the PCA model, and the Q statistic was calculated for each process measurement in the training data set. In this sense, Figure 1 shows the Q statistic for each observation sampled every 20 h for all the simulated photo-fermentations. The Q statistic threshold was calculated with 95% confidence level.

In general, PCA results showed that observations below the control limit (Q statistic threshold) were those belonging to process simulations under light intensities between 100 and 340 W/m² and pH initial values between 6.3 and 7.5, regardless of some exceptions. Therefore, the normal or zero class was defined under these process conditions, and the rest of the conditions were classified into four classes of faults considered as undesired scenarios of the photo-fermentation process, which are LI < 100 W/m², LI > 340 W/m², pH < 6.3, pH > 7.5, named as classes 1 to 4 in that order, as shown in Table 1.

The whole data were arranged in the order of these classes into the training set, and a classifier or label matrix $y_i \in \{+1, -1\}$ with four columns as the number of classes was created. SVM were then applied to both training data set and the classifier matrix to develop four classifiers, which were the data-based diagnosis models. A linear kernel function in the SVM method, which converts a non-linear classification problem into a linear one with a highly dimensional feature space by relating w, x and b parameters with a linear function, turned out to be the best for learning and classifying the data into the four classes.

Once the data-based diagnosis models were constructed in Matlab software, these models were tested on a data set called the validation set, which was formed of simulation runs different to those included in the training set in order to assess the diagnosis performance of classification models in terms of the well known F1 score (Monroy et al. 2022a, 2022b), which is calculated as a weighted average of both precision and recall or sensitivity terms, according to Equation (2).

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision measures the proportion of assigned observations to the positive class with respect to the total number of predicted positive outcomes to one class while Recall measures the ratio between the number of positive outcomes that are correctly identified and diagnosed with respect to the total number of real positive observations from a class.

In this context, Figure 2 shows the diagnosis or classification performance in terms of the F1 score, obtained after the data models were evaluated on the validation data set (blue bars).

In addition, diagnosis models were tested on some laboratory experimental photo-fermentations (red bars) to demonstrate their validity and generalization capacity. As observed in Figure 2, a 100% diagnosis performance was achieved in process scenarios where light intensities were either lower or higher than the desired range, in both simulated and experimental photo-fermentations. On the other hand, the diagnosis performance was lower for those photo-fermentations under pH values out of the nominal range.

Nevertheless, it is important to highlight two important things. First, no experimental photo-fermentations with initial pH < 6.3 were disposed, which explains the result obtained for class 3 in the experimental data set (zero performance). Second, a deep analysis of the model predictions for classes 3 and 4 (blue bars) showed a missed diagnosis to the zero class during the first observations of those photo-fermentations simulated with the mechanistic process conditions.

### Table 1 | Photo-fermentation process conditions (pH and light intensity) for the classes of both desired and undesired scenarios and diagnosis models

<table>
<thead>
<tr>
<th>Class of process scenario (normal and undesired)</th>
<th>Process conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 0</td>
<td>LI $\leq$ 340 W/m², pH $\leq$ 7.5</td>
</tr>
<tr>
<td>Class 1</td>
<td>LI $&gt;$ 340 W/m², pH $\leq$ 7.5</td>
</tr>
<tr>
<td>Class 2</td>
<td>LI $\leq$ 100 W/m², pH $&gt;$ 7.5</td>
</tr>
<tr>
<td>Class 3</td>
<td>LI $\leq$ 340 W/m², pH $&gt;$ 7.5</td>
</tr>
<tr>
<td>Class 4</td>
<td>LI $\leq$ 100 W/m², pH $&gt;$ 7.5</td>
</tr>
</tbody>
</table>

Figure 1 | Q statistic for all the observations in the training data set, obtained from simulation runs with the mechanistic model.
model (mainly for batches run under pH < 6.3). However, after some simulation hours, the diagnosis was right which evidenced that the photo-fermentation process is very sensitive to pH changes as it has been revealed in previous research (Monroy et al. 2016), and that pH oscillations away from the optimum range are detected with delay.

In general, results obtained from evaluating the diagnosis models on experimental data supported their validity and evidenced their potential to be used in future applications related to monitoring those process variables that are harsh to measure or quantify on-line such as the LI.

CONCLUSIONS

Data-based diagnosis or classification models were constructed using several simulated batches from the hydrogen production by photo-fermentation and applying the SVM method. Classification models were based on supervised learning, which consists of labelling all the experiments and batches according to classes of faulty or undesired scenarios so that the classification method can learn and train from these data. The resulting models are then expected to detect and diagnose those scenarios in further experiments of the process.

Process data were obtained from simulations with a mechanistic model previously developed in Matlab and reported for the photo-fermentation process carried out by an immobilized consortium of photo-bacteria. Four diagnosis models were produced as the number of undesired scenarios found in the process data, revealed by Q statistic values over a threshold (Q limit), both calculated after applying PCA to a training data set of the process.

Diagnosis models were validated on different batches from the same process, not only simulated, but also with experimental photo-fermentations, obtaining a correct diagnosis of the undesired process scenarios, given by high diagnosis performance indices (F1 score), which evidenced the reliability of the models to be further applied on detecting and diagnosing LI and pH values out of optimal ranges in outdoor photo-fermentations.

ACKNOWLEDGEMENTS

This research work was supported by CONACYT-Ciencia Básica [251718]. Financial support from CONACYT through the SNI program is fully appreciated. We highly acknowledge Eliane Guevara López for her experimental work and Jaime Pérez for his technical assistance in the laboratory. Finally, this paper would not have been published without the support provided by the programme committee of the International Young Water Professionals Conferences organized by IWA, WISA and YWP-ZA.

REFERENCES

Ardakani, M. H., Askarian, M., Escudero, G., Graells, M. & Espuña, A. 2017 Toward online explore of concept drift for


