Computers and Chemical Engineering 194 (2025) 108991

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Towards a machine learning operations (MLOps) soft sensor for real-time predictions in industrial-scale fed-batch fermentation

Brett Metcalfe [a,b], Juan Camilo Acosta-Pavas [c], Carlos Eduardo Robles-Rodriguez [c], George K. Georgakilas [d], Theodore Dalamagas [d], Cesar Arturo Aceves-Lara [c], Fayza Daboussi [e], Jasper J Koehorst [a,f], David Camilo Corrales [c,*]

[a] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands
[b] Department of Earth Sciences, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[c] TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France
[d] Information Management Systems Institute (IMSI), ATHENA Research Center, Athens 15125, Greece
[e] INRAE, UMS (1337) TWB, 135 Avenue de Rangueil, Toulouse 31077, France
[f] UNLOCK Large Scale Infrastructure for Microbial Communities, Wageningen University & Research, Gelderland, Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Machine learning pipelines
MLOps
LTSM
Soft-sensors
Concept drift
IndPenSim

ABSTRACT

Real-time predictions in fermentation processes are crucial because they enable continuous monitoring and control of bioprocessing. However, the availability of online measurements is limited by the availability and feasibility of sensing technology. Soft sensors - or software sensors that convert available measurements into measurements of interest (product yield, quality, etc.) - have the potential to improve efficiency and product quality. Machine learning (ML) based soft sensors have gained increased popularity over the years since they can incorporate variables that are measured in real-time, and exploit the intricate patterns embedded in such voluminous datasets. However, ML-based soft sensor requires more than just a classical ML learner with an unseen test set to evaluate the quality prediction of the model. When a ML model is deployed in production, its performance can deteriorate rapidly leading to an unanticipated decline in the quality of the output and predictions. Here a proof concept of Machine Learning Operations (MLOps) to automate the end-to-end soft sensor lifecycle in industrial scale fed-batch fermentation, from development and deployment to maintenance and monitoring is proposed. Using the industrial-scale penicillin fermentation (*IndPenSim)* dataset that includes 100 fermentation batches, to build a soft sensor based on Long Short Term Memory (LSTM) for penicillin concentration prediction. The batches containing deviations in the processes (91–100) were used to assess concept drift of the LSTM soft sensor. The evaluation of concept drift is evidenced by the soft sensor performance falling below the set threshold based on the Population Stability Index (PSI), which automatically triggers an alert to run the retraining pipeline.

1. Introduction

1.1. Industrial biotechnology

Industrial biotechnology (IB) - the production of products or intermediary products via the biological pathways of (un)modified organisms - can for example be used in the production of food (*e.g.*, food additives, drinks), biofuels (Nielsen et al., 2013; Clomburg et al., 2017), chemicals (Meadows et al., 2016; Clomburg et al., 2017), and biopharmaceuticals (*e.g.*, drugs, vaccines) (J. Zhang et al., 2022; Naseri, 2023) via large-scale fermentation (Erickson, 2009). Besides the production of new and novel chemicals IB as an approach could be used to circumvent traditional, unsustainable, carbon intensify production practices by replacing them with sustainable alternatives (Nielsen et al., 2013; Naseri, 2023) helping to transition to a circular (Meyer et al., 2020) and / or green economy (Shayegh et al., 2023).

A common method in IB is the growth of individual, or populations of, organism(s) within a fermenter to produce a product. With the fermentation procedures being typically described by a set of variables that correspond to distinct quantitative and qualitative features; *i.e.,* temperature, pH, gas concentrations, feed rate, fermentation phase, product concentration *etc*. The product yield from this method is

* Corresponding author.
E-mail address: David-Camilo.Corrales-Munoz@inrae.fr (D.C. Corrales).

https://doi.org/10.1016/j.compchemeng.2024.108991
Received 20 September 2024; Received in revised form 2 December 2024; Accepted 23 December 2024
Available online 26 December 2024

dependent upon, or controlled by, certain abiotic (*e.g.*, temperature, pH) and biotic factors referred to as critical process parameters (CPP). These CPP may have a negative, or a positive, impact on the desired product and may directly or indirectly impact its Critical Quality Attributes (CQA) (Cardillo et al., 2021; Smiatek et al., 2020). Furthermore, a lack of comprehension regarding the influence of diverse parameters on biological processes can result in confounding effects, thereby impeding the ability to anticipate the consequences of process alterations. This contrasts with the greater certainty afforded by more traditional manufacturing processes, where input-output relationships are more constrained and predictable. Collectively, these factors restrict the capacity to directly steer the fermentation process, thereby limiting the efficacy of IB.

Some of these variables can be measured continuously on frequent time intervals (real-time/online) by fermentation equipment sensors and actuators, while others can only be measured offline due to technological limitations, after extracting samples from the fermentation vessel during or after the procedure finishes. In addition to online and offline monitoring, other strategies such as inline and at-line monitoring are becoming increasingly significant within the field of process analytical technology (PAT). Inline monitoring refers to the continuous measurement of process variables directly within the fermentation stream, thereby providing real-time data without the necessity for sample extraction (Claßen et al., 2017). In contrast, at-line monitoring entails the measurement of samples obtained at designated points in the process, which are then analysed in close proximity to the process itself (Vojinović et al., 2006). This approach effectively reduces the time delay inherent to traditional offline analysis. The integration of these methods, in conjunction with advancements in sensor technology and real-time analytics, has markedly enhanced the capacity to monitor and regulate fermentation processes with greater efficiency. However, the implementation of PAT is still limited by the lack of devices designed for specific molecules of interest, product quality, or yield in real time within the bioreactor.

1.2. Soft sensors

With the potential long development times of PAT devices - which cannot be readily circumvented - significant effort has therefore, been placed on using the abundant data used for monitoring the fermentation to provide a continuous estimation of the process quality and product concentration in real time through modelling (Helleckes et al., 2023; Lawrence et al., 2024). Such models are called soft sensors, which have emerged in the process industry as a solution to provide online estimations to other process variables which are measured either indirectly or at ineffective rates (Ji et al., 2023). The word “soft sensor” is a combination of “software” given that the models are usually computer programs, and “sensors” since the models provide similar information as physical sensors (Kadlec et al., 2009). The soft sensor modelling approaches can be classified into model-based, data-driven (Fortuna et al., 2007), and hybrid. Model-based soft sensors are based on the first principle of physical or chemical laws to provide detailed process description. Whereas data-driven soft sensors are created based on large datasets using machine learning (ML) techniques (Yan et al., 2017). ML-based soft sensors have gained increased popularity over the years since they can incorporate variables that are measured in real-time, and exploit the intricate patterns embedded in such voluminous datasets. In addition to parameter estimation, soft sensors offer the opportunity to develop sensors for other quantitative and qualitative uses, for instance in IB sensing change during the switch-over from one fermentation phase to the other; testing different strains under the same conditions; or benchmarking different conditions using the same strain. Furthermore, soft sensors can give useful information in terms of fault detection by working with sensors in parallel (Huang et al., 2002; Kaneko et al., 2009) providing real-time monitoring of fermentation processes that are robust to any type of deviation from the norm.

1.3. Soft sensors in bioproduction

Recently, several papers have focussed upon the development of soft sensors for bioprocessing. For instance, an adaptive soft sensor solution for biomass monitoring was proposed by (Siegl et al., 2022). This approach is based on distinct submodels that operate on different sensor or actuator data that can be used to model *Pichia pastoris* biomass concentration, such as acids and nitrogen consumption, gas measurements, metabolites and media composition (Siegl et al., 2022). The various submodels were combined using a real-time ensemble strategy which estimates the reliability of each submodel on a moving window regression approach. To incorporate the adaptive component, each submodel is penalised by comparing current and previous prediction using a *t*-test, and the final prediction calculated after applying variance-based weighting. In contrast, a different approach toward developing a soft sensor was adopted by (Siegl et al., 2023) that promises to adaptively and automatically recalibrate the underlying modelling procedure to accommodate accurate predictions in diverse bioprocesses. To achieve these results the algorithm utilises multilinear principal component analysis (PCA) to calculate an euclidean-based distance matrix as input for k-nearest neighbour (kNN) clustering. The resultant clustering is then used to select the historical data that will be used for the recalibration process, which will subsequently be used to also detect phase switching time points. This framework has been applied to estimate the biomass in a *P. pastoris* process as well as the biomass and protein concentration in a *Bacillus subtilis* fermentation process described by variable conditions.

In comparison, a more versatile approach was presented in (Qiu et al., 2020) that exploits semi-supervised learning to model multiphase batch processes using unlabelled data from simulation platform of the penicillin fermentation process named PenSim (Birol et al., 2002). The proposed framework uses an adversarial autoencoder to assign labels to learn from the process mechanism both the spatial and sequential distances to estimate the labels of the abundant unlabelled data that are derived from sensors. A Gaussian mixture model is applied to distinguish between different batch phases and the just-in-time relevance vector regression model estimates the target variable. Similarly, (Qiu et al., 2022) also proposed the development of a semi-supervised approach based on relevance vector machine algorithm (Tipping, 2001) for the analysis of multiphase batch processes in PenSim (Birol et al., 2002). The initial stage of their proposed methodology involves the utilisation of a sequence-constrained fuzzy c-means algorithm, which is employed to partition asymmetric data into distinct phases (Qiu et al., 2022). Subsequently, a localised semi-supervised learning (LSL) algorithm is introduced to estimate labels for the unlabelled data in each phase. The LSL algorithm comprises two main steps: the construction of a similar dataset based on a designed comprehensive similarity measure followed by label estimation using the Just-In-Time Learning (JITL) algorithm.

More recently, the development of interpretable soft sensors for monitoring and controlling the penicillin concentration was explored (Acosta-Pavas et al., 2024) using an industrial penicillin simulation dataset named IndPenSim (Goldrick et al., 2015, 2019). Variables such as substrate feed rate, agitation, temperature, pH, dissolved oxygen, vessel volume, CO_2, and O_2 percent in off-gas were considered as predictors for the production of penicillin within the fermentation using interpretable learners. Where, Interpretable learners are founded on a set of transparent principles, comprising a tree-based structure, rule-based representation, an analysis of feature importance, and a focus on model simplicity (Acosta-Pavas et al., 2024). This work compares interpretable learners such as Classification And Regression Tree (CART) (Breiman et al., 1984; Stulp and Sigaud, 2015), M5 (Quinlan, 1992; Wang and Witten, 1997), CUBIST, and Random Forest.

1.4. Aims and objectives

As soft sensors gain traction in bioprocessing (*e.g.*, (Acosta-Pavas

et al., 2024)) there must be a transition from their testing and validation on post fermentation datasets to their usage and operational deployment during fermentation. To utilise a soft sensor in IB 'production' requires more than just a classical ML model with an unseen test set to evaluate the quality prediction of the model; it demands a comprehensive architecture that is deployable within an IB setting. Furthermore, in order to fully realise the potential of soft sensors in IB, it is essential that their end-to-end lifecycle be automated. Such automation enables scalability, ensures consistency, and provides the requisite adaptability for fermentation processes. In contrast to traditional soft sensors, which are evaluated on static datasets, soft sensors in IB production must continuously adapt to changing conditions, such as variations in substrate or operational parameters (Siegl et al., 2023). Such an architecture will be a machine learning operations (MLOps) in industrial fed-batch scale fermentation. Where, MLOps represents a set of practices, tools, and methodologies that are designed with the objective of streamlining and automating the deployment, monitoring, and life cycle management of machine learning models in production environments (Treveil et al., 2020). The automation of processes, including model retraining, deployment, and monitoring, through MLOps practices not only reduces operational costs but also enhances reliability and delivers real-time insights critical for process optimisation. These advancements highlight the transition from experimental validation to practical deployment, underscoring the importance of integrating soft sensors into fully operational architectures tailored for IB environments.

Here a proof of concept for a bioprocessing MLOps pipeline using a soft sensor is outlined in order to investigate and assess what components are necessary (*e.g.*, a data producer, a time series database) as well as what concepts need to be taken into account (*e.g.*, concept drift) and/or adjusted (*e.g.*, data pre-processing, training and retraining). The aim of this paper is to assess the feasibility of machine learning pipelines using a 'perfect' dataset, *i.e.*, a high quality, well structured, dataset that incorporates various levels of aggregation. The manuscript is organised as follows: first a description of the components and processes of the MLOps pipeline will be presented as well as how they were simulated in this proof of concept. This description is followed by an explanation of the soft sensor based on Long Short-Term Memory network (LSTM) for penicillin concentration prediction that has been developed as well as how it has been validated. Following this, the concepts introduced in the MLOps are introduced and discussed. Finally, the manuscript will end with a future outlook.

2. MLOps pipeline for real-time predictions

2.1. MLOps pipeline architecture

The MLOps pipeline architecture for the soft sensor has been designed with the objective of facilitating the seamless integration, deployment and ongoing maintenance of machine learning models. This pipeline automates the soft sensor's lifecycle, including data ingestion, model training, validation, and deployment. The pipeline typically incorporates real-time data collection, whereby sensor and actuator inputs such as temperature, pH and nutrient levels are fed into the model for prediction. Following the training and evaluation of the model, the system is able to automatically deploy updated models, thereby reducing downtime and the necessity for human intervention.

Architecturally our pipeline is divided into components within the physical world (*e.g.*, sensors, actuators, control systems) that may produce continuous data streams (*e.g.*, sensors) or discrete datasets (*e.g.*, sampling) and those in the virtual world (Fig. 1). The virtual world is itself subdivided into computational (*e.g.*, time series database, processing scripts, model scripts, *etc.*) and visualisation components (*e.g.*, the dashboard section presented in *supplementary material*). In the following sections the various aspects, components, and modules of a MLOps pipeline are presented. Whilst this architecture has been designed for the specific context of real-time predictions within the domain of bioprocessing it is potentially applicable for other use cases (*e.g.*, agricultural greenhouse; farming; *etc.*).

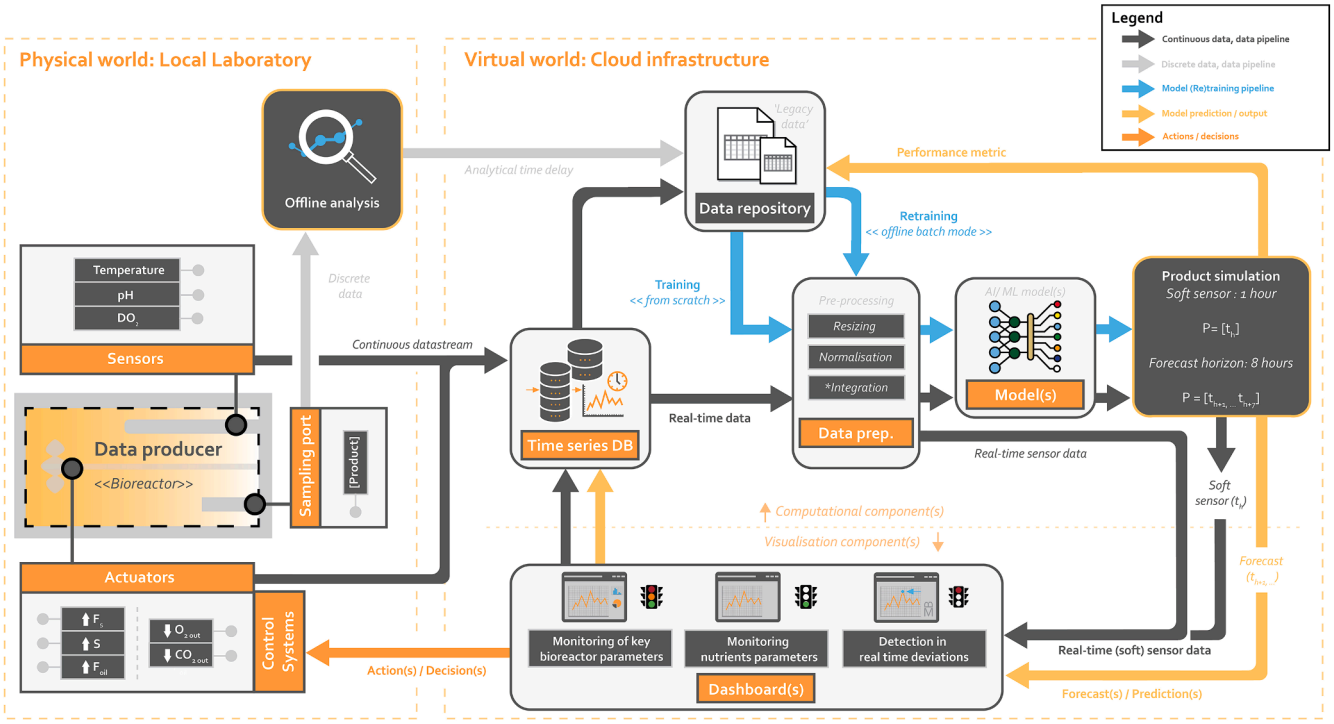

Fig. 1. Architecture of the MLOps pipeline for real-time predictions of soft sensors in bioprocess. The laboratory in the physical world is coupled to a virtual world divided between computational components (*i.e.*, processing and modelling steps) and visualisation components (*e.g.*, dashboards). Data produced by sensors, actuators, and sampling (see Figure A4 in supplementary material) used for the soft sensors are described in Table 1.

2.2. Data producer

2.2.1. Data producer: fermentation vessel

A data producer is the root source of data for the MLOps pipeline which for a IB pipeline will be a bioreactor or fermentation vessel. A bioreactor is a vessel or tank in which whole cells or cell-free enzymes transform raw materials into biochemical products and/or less undesirable by-products (Erickson, 2009). In a classical bioreactor, key variables need to be controlled to ensure that cells or enzymes are maintained at the desired environmental conditions to achieve a specific product at the desired yields or productivities and within the desired process times.

Certain bioreactors may have combined control systems for these variables that permit the adjustment of conditions within the bioreactor based on input from various integrated sensors that monitor variables like temperature, pH, dissolved oxygen, agitation, nutrient addition and biomass concentration. These sensors provide online measurements for process control and optimization of bioprocesses.

Components can be:

- Agitator or stirrer: the agitator is a mechanical component that mixes the contents of the vessel. It ensures uniform distribution of nutrients, oxygen, and temperature throughout the culture. Common types include impellers and paddles.
- Heat exchanger: a heat exchanger is used to control the temperature inside the bioreactor. It applies either a heating or cooling to the culture medium as needed to maintain the desired temperature for cell growth or fermentation.
- pH: a method by which to regulating the culture's pH through the addition of an acid or base, it is essential for cell viability and optimal product yield.
- Aeration and gas supply: either a gas supply or method by which to aerate the mix in order to maintain the necessary oxygen levels in the medium to support desired cell growth.
- (Nutrient) Inlet-outlet: an inlet that delivers nutrients, such as sugars, salts, vitamins, and trace elements, to the bioreactor. Likewise, waste products are removed through an outlet to maintain a balanced culture medium.

During cultivation, off-line measurements can be taken through sampling ports, allowing samples of the culture to be collected for analysis without disturbing or disrupting the process. Sampling is essential to assess cell density, product concentration, and other parameters via offline measurements. Data generated by a bioreactor can be characterised as pertaining to one of three components: sensors, actuators or sampling port (Table 1).

Table 1
Summary of the key bioreactor parameters of the IndPenSim dataset used in this study. *Note in the dataset used the Ammonia shots have only the value 0 throughout the dataset.

Parameter	Bioreactor component	Description	Units
T	Sensor	Temperature of broth	K
pH	Sensor	Acidity of an aqueous solution	–
DO_2	Sensor	Dissolved oxygen concentration	mg L^{-1}
F_s	Actuator (Nutrient inlet)	Sugar flow rate	L h^{-1}
S	Actuator (Nutrient inlet)	Substrate concentration	g L^{-1}
F_{oil}	Actuator (Nutrient inlet)	Soybean oil flow rate	L h^{-1}
NH_3_shots	Actuator (Nutrient inlet)	Ammonia shots*	kg
O_{2out}	Sensor (Nutrient outlet)	Oxygen off-gas concentration	%
CO_{2out}	Sensor (Nutrient outlet)	Carbon dioxide off-gas concentration	%
P	Sampling port	Offline penicillin concentration	g L^{-1}

2.2.2. Data producer: indpensim

The dataset used in this work corresponds to a series of experiments that were subsequently utilized to produce a simulation of industrial-scale penicillin fermentation processes (Paul and Thomas, 1996) which describes all the component balances relating to the process variables (Goldrick et al., 2015, 2019). The dataset generated from the industrial-scale penicillin fermentation (*IndPenSim)* includes 100 batches with each dataset comprising 2238 variables of which 39 variables correspond to process variables (manual and automatic control and online and offline measurements) and the remaining 2199 correspond to Raman spectra. Per batch sensors recorded data every 12 min with the average batch length being approximately 230 h. Out of the 100 batches, the first 90 batches were operated under "normal" conditions using three different control strategies: (I) controlled by a recipe driven approach; (II) controlled by an operator and (III) controlled by an Advanced Process Control (APC). The last 10 batches contains faults, resulting in process deviations (Goldrick et al., 2019). These batches are further expanded in what follows:

I. Controlled by recipe driven approach (IndPenSim batches 1–30): recipe driven manipulation of certain variables such as sugar feed rate and phenylacetic acid flow rate. For the soft sensor training it replicates a procedure driven by a particular set of rules. A fixed profile was used throughout the batch.
II. Controlled by operators (IndPenSim batches 31–60): an operator manipulates the fixed profile throughout the batch (*i.e.*, operator dependent) replicating for the soft sensor training the manual actions performed by the operators such as adjusting sugar flow rate (Fs) and phenylacetic acid (PAA) throughout the batch.
III. Controlled by an Advanced Process Control - APC (IndPenSim batches 61–90): a PAT analyser manipulates the fixed profile throughout the batch. These batches used an empirical mathematical model to simulate a realistic PAT analyser (Raman spectroscopy device).
IV. Batches with faults resulting in process deviations (IndPenSim batches 91–100): the final batches represent those with faults allowing the soft sensor to train on 'realistic' data rather than perfect datasets. The IndPenSim model was used to simulate faults in aeration, pH sensor drifts, variations of substrate concentration and coolant by implementing two standardised multivariate data analysis (MVDA) as fault detection algorithms.

The strain studied in this bioprocess was the *Penicillium chrysogenum* run in a 100,000-litre vessel with a radius of 2.1 m. The vessel was equipped with multiple sensors and three Rushton impellers (radius of 0.85 m) that are operated at a fixed agitation speed of 100 rpm (see Figure A4 in supplementary material)

2.3. Simulated data producer

In this paper, only a subset of the data generated by the bioreactor has been considered. To mimic a real fermentation using a network-capable fermentor, the 100 batch IndPenSim dataset (Goldrick, 2019; Goldrick et al., 2019) was downloaded and subdivided into individual batch specific files with a subset of the available data. From the sensors the temperature, pH and dissolved oxygen concentration parameters were selected whilst from the nutrient inlet-outlet variables included the soybean oil flow, ammonia shots, sugar flow rate and substrate concentration. In the dataset, soybean oil was used as both a secondary carbon source as well as acting as an anti-foaming agent whilst low levels of nitrogen were rectified through the addition of ammonia sulphate shots (Goldrick et al., 2019).

The parameter 'NH_3_shots' (Table 1) only had values equal to 0, in other words, ammonia shots were not injected. The data associated with

nutrient inlets have a stair, step or plateau type of pattern. This means that the nutrient inlets remain fixed during a long period, or they do not change over time. The remaining variables in the bioreactor, such as temperature, pH, dissolved oxygen, and nutrient outlets, correspond to continuous data. The penicillin concentration used as the target variable for the soft sensor is a discrete (although recorded in the dataset as a continuous stream of data) dataset from bioreactor samples which are analysed in a laboratory, and therefore is an off-line measurement.

The variables selected in this study were chosen to represent the key parameters of the bioreactor, which were categorised as sensors, actuators and controllers based on the input of domain experts and their relevance for real-time data. Sensors provide real-time measurements of critical environmental and process conditions, including temperature (T), acidity (pH), dissolved oxygen (DO_2), and gas concentrations (O_{2out}, CO_{2out}). Actuators, conversely, reflect controllable actions to the system, including nutrient flow rates (F_s, S, F_{oil}) and ammonia shots (NH_3_shots).

2.4. Data repository

Once a batch has been completed the various datasets must be stored for future use. Here, the term "legacy data" is used to describe existing 'historical' data that has been collected previously by the "data producer" from its actuators, sensors and sampling ports (Fig. 1) and/or data that may originate from previous experiments including from different types of bioreactors (*e.g.*, brand, volume, *etc*.). Or from the same type of bioreactor but under different experiment conditions (*e.g.*, environmental, strain, organism, *etc*.). Legacy data enables the initial testing and training of the (soft sensor) model.

2.5. Data preparation

The data preparation module is responsible for cleaning and processing and can be used in either a real-time (Section 2.3) or historic legacy data mode (Section 2.4). A data preparation model ensures that the data used to train the soft sensor is clean, relevant and formatted correctly as frequently legacy data will be generated in various user, laboratory, or institutional specific formats, including but not limited to relational/non-relational databases, excel spreadsheets, csv, json, etc. All of which may, or may not differ from the original format of the data producer. In the context of ML regression models, data quality issues such as missing values, outliers, redundancy, and noise need to be carefully addressed depending on the nature of the problem (Corrales et al., 2018, 2020). However, in the present study, the penicillin data originate from a controlled mechanistic model, and thus, these data quality issues are not a concern. For this specific case of penicillin fermentation, we will address techniques for data preprocessing, including resizing, data integration, and feature scaling, which will be presented subsequently.

2.5.1. Data integration between online and offline measurements

Data captured from sensors and data captured from sampling ports are not aligned perfectly in time. For example, the data of the sampling ports are offline measurements that could be analysed in the laboratory many days after sample collection, and they may not correspond exactly to the timestamps of sensor readings. The data are not in a one-to-one correspondence, and this type of process data is referred to as asymmetric data (Qiu et al., 2022). If the sampling rates of the online and offline measurements differ, the data with the lower sampling rate should be interpolated, aggregated to a common time resolution, or utilised to create a dynamic model to increase the volume of data.

In the case study presented in this paper (IndPenSim), the offline measurements of penicillin concentration were symmetrically aligned with the online measurements, thanks to the industrial-scale penicillin model (Paul and Thomas, 1996) adapted by (Goldrick et al., 2015). The dynamic model accounts for the growth, morphology, metabolic production, and degeneration of biomass during a submerged *P. chrysogenum* fermentation. More detailed information about the dynamic model can be found in Goldrick et al. (2015). The dynamic model generates five points per hour, but these points are not assigned to specific times. For the purposes of this study, it is assumed that each point within an hour represents a measurement taken every 12 min.

2.5.2. Resizing

The resizing module is utilized to dynamically prepare the training set and input data of the model in real-time. Resizing refers to the process of changing the dimensions, or size, of the data with the purpose to prepare for integration (Section 2.5.1) or use of the ML soft sensor. In this context, the data is prepared as a time series using a data window technique (Harris, 1978) that consists of creating a time window for all predictors in the dataset into chunks of M consecutive observations thereby creating a new dataset with fewer rows but retaining the same number of columns. Furthermore, this new dataset shares data chunks between adjacent windows, for example, input data at a timestamp of 31 hr will share data from timestamps 29 and 30 h with the adjacent input data at time 30hr (see Fig. 2). The data window technique considers two parts:

- Input window: a data sequence with a certain number of past time chunks. For ML soft sensor, the size of the input window is set for 4 time chunks:

  Input Window: $[X_{(t-3)}, X_{(t-2)}, X_{(t-1)}, X_{(t)}]$
  Where X corresponds to input features:
  X: [T, pH, DO_2, F_s, S, F_{oil}, O_{2out}, CO_{2out}]

- Forecast horizon: data points in the following sequence or future time chunks of penicillin concentration. The output target *(Y)* is a sequence of future values of penicillin concentration *(P)* over a specific forecast horizon. For ML soft sensor, the forecast horizon is the future 8 hrs:

  Y: $[P_{(t)}, P_{(t+1)}, P_{(t+2)}, P_{(t+3)}, P_{(t+4)}, P_{(t+5)}, P_{(t+6)}, P_{(t+7)}]$

To illustrate the dataset preparation of the resizing module, consider a data sequence with an input window of three-time chunks and a forecast horizon of eight-time chunks (Fig. 2). The input window comprises the parameters of the bioreactor data from hours 28, 29, and 30, while the forecast horizon encompasses the penicillin concentration from hours 30, 31, 32, 33, 34, 35, 36 and 37. This implies that the ML soft sensor will utilise data from hours 28, 29, and 30 to predict values for hours 30 through to 37.

2.5.3. Feature scaling

The next step of the data preparation is feature-scaling which has the aim to transform the variables of the dataset onto a consistent scale. For bioprocessing, particularly for data coming from various sensors connected to a bioreactor, this step is a necessity as the variables will have different ranges, units of measurement, or orders of magnitude. For example, temperature is expressed in Kelvin [K], dissolved oxygen concentration is measured in milligrams per litre [mg L^{-1}], and sugar flow rate represented in litres per hour [L h^{-1}] (Table 1). Too much variation between variables can lead to biased model performance or difficulties during the learning process. Therefore, the *IndPenSim* dataset was standardised using a Z-score:

$$z = \frac{x - \mu}{\sigma}$$

Where μ and σ is the mean and the standard deviation of the training samples respectively and x is the sample to standardise. The Z-score is useful when a dataset distribution is Gaussian or unknown (*i.e.*, assumed normality); in addition, it is less sensitive to outliers and preserves the relationships between the data points (Saisana, 2014).

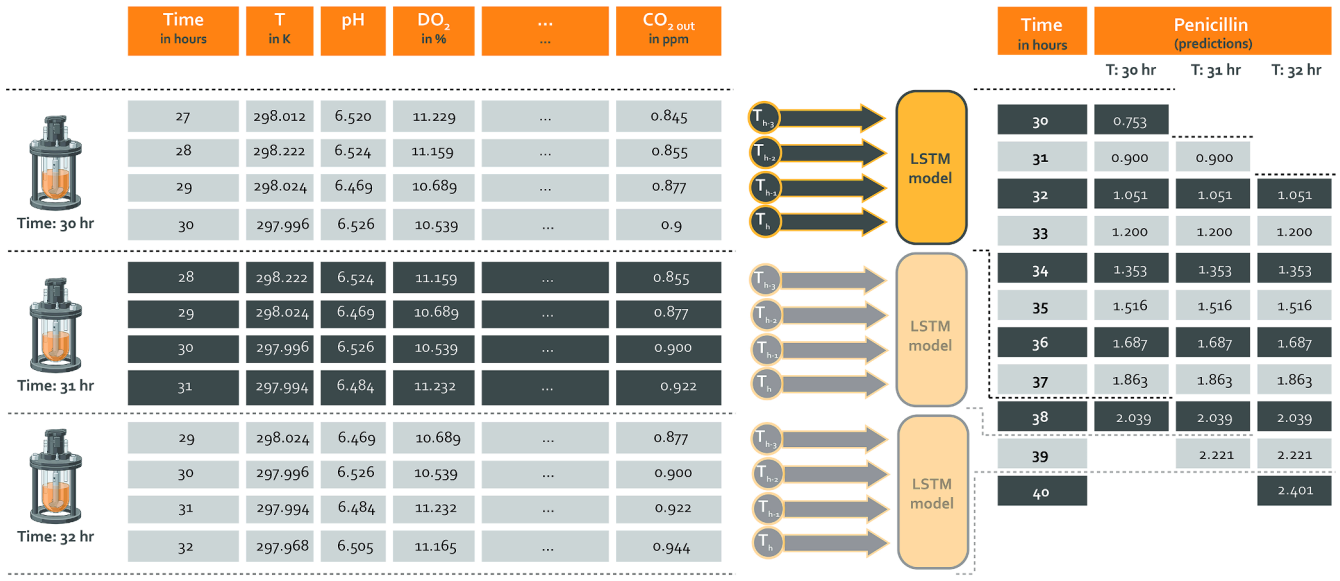

Fig. 2. Data sequence preparation for ML soft sensor. On the left the input data with an input window of 4 time chunks and on the right the forecast horizon of 8 time chunks. For example, at time 30 hr the input data is 4 h (*i.e.*, the data for that hour and 3 h preceding so between 27 and 30 h) and the output data is 8 h ahead (*i.e.*, between 30 and 37 h). The selection of time windows was performed empirically by testing window sizes between 3 and 8, with the optimal results obtained at a window size = 4. However, this time window may vary depending on the data distribution, which is linked to the experimental conditions, the type of bioreactor, and the microorganism used.

2.6. Machine learning model

2.6.1. Soft sensor for prediction of penicillin production

Once the data has been prepared, the machine learning (ML) model is either (re-)trained upon it, validated against it or used as input. Although the definition of a soft sensor involves simulating an online measurement, this study aims to emphasise the significance of not only providing an estimated online value but also having a forecast horizon of penicillin production (Section 2.5.2) to facilitate decision-making in the experiment. The soft sensor proposed is based on recurrent neural networks (RNNs) for addressing time-dependent problems (Elman, 1990; Hochreiter and Schmidhuber, 1997). RNNs can capture and retain temporal dependencies, even across extended periods (Mandis et al., 2024). Among RNNs algorithms, we used Long Short-Term Memory network (LSTM) to build the soft sensor which is specifically designed to overcome the vanishing gradient issue by leveraging its specialized units.

LSTM units consist of memory cells and three nonlinear gates for regulating the flow of information: the input gate, forget gate and output gate (Fig. 3), where h_{t-1} and h_t denote the outputs of LSTM at the previous and current moment, respectively. Similarly, c_{t-1} and c_t represent the states of the memory units (cell states) at the previous and current moment. The input is represented by x_t.

The implementation of the three gates is given by the following equations (B. Wang et al., 2023):

Forget gate: f_t determines how much of the t-1 moment cell state is retained in the current cell state:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$$

Input gate: i_t determines how much of the newly obtained informa-

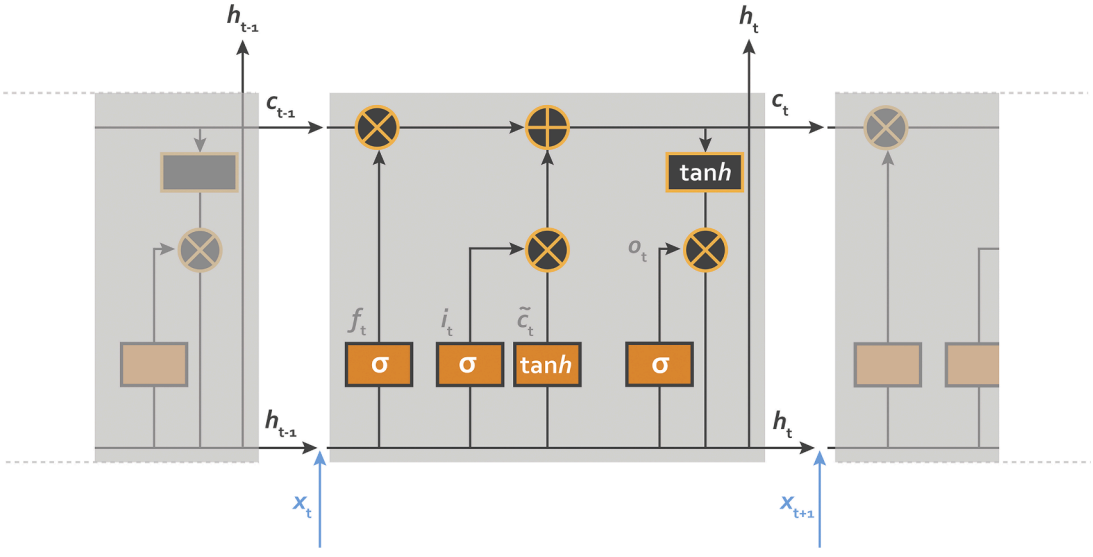

Fig. 3. Schematic of the structure of a series of LSTM cells. Structure of LSTM cell where x_t is the input data, h_t is the output data, and c_t is the storage unit state at time t. Within the cell f_t, i_t, o_t and $\tilde{}c_t$ are forget gate, input gate, output gate, and newly obtained information respectively. Adapted from (B. Wang et al., 2023).

tion c_t needs to be updated, and the calculation results become part of the updated neuronal state c_t:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$\widetilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c)$$
$$c_t = f_t c_{t-1} + i_t \widetilde{c}_t$$

Output gate: o_t determines how much information in the updated neuron state c_t becomes a hidden layer state variable h_t:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t tanh(c_t)$$

Where: W_f, W_i, W_c and W_o are the weight matrices of the corresponding control gates;b_f, b_i and b_o are the corresponding bias vectors; σ is a Sigmoid activation function with a value range of [0,1], 0 for all discarded, 1 for all reserved; *tanh* is a hyperbolic tangent activation function, $\odot$ represents the product of the Hadamard.

2.6.2. Hyperparameters selection for LSTM soft sensor

The LSTM network proposed in this study was implemented using Python 3.9.14, leveraging the Keras 3.5.0 library with Tensorflow 2.17 as its backend framework. The appropriate architecture for the LSTM network, central to the soft sensor model, involves the selection of the suitable values of the hyperparameters during the training phase. The optimal hyperparameters of the LSTM soft sensor were determined through systematic experimentation, varying the hyperparameters of Table 2. Each configuration undergoes evaluation using a loss function to measure the error. Following experimentation, hyperparameters of the LSTM network architecture underlined in Table 2 demonstrated the highest levels of training and validation accuracy.

In this instance, the LSTM model was trained with 13,742 time points corresponding to the full datasets of the following IndPenSim fermentation batches: control strategies by recipe driven approach (IndPenSim batches 1–30) and by operators (IndPenSim batches 31–60). Batches with process deviations (IndPenSim batches 91–100) were used to evaluate and retrain the LSTM model. The LSTM soft sensor was constructed using a 10-fold cross-validation with three repetitions. The data for each batch is exclusively assigned to either a training fold or a validation fold. In this study, with 60 batches and 10 folds, each fold contained precisely 6 batches. Figure A*4*In supplementary material, Figures A1, A2 and A3 show the validations of the LSTM model by fold and repetition.

Although all hyperparameters presented in Table 2 are important for ML model construction, the choice of batch size was the most critical. The batch size is a number of samples processed before the LSTM weights are updated. We tested three different batch sizes (1, 250, 13,742) following the definition of the gradient descent methods (Goodfellow et al., 2016). In batch gradient descent, all training instances (13,742) are used to create one batch. Whereas the batch size is equal to one sample, the gradient descent method is called stochastic. However, when the batch size is greater than one sample but less than the size of the training dataset (250), the gradient descent is called mini-batch. We obtained the best results with mini-batch gradient descent, where we defined a batch size equal to 250, reflecting the number of time steps in each fermentation batch (approximately 250 h). This implies that each LSTM batch contains all instances from a single fermentation batch.

Table 2
Hyperparameters selected for LSTM soft sensor. The text highlighted in bold corresponds to the hyperparameters used for the construction of the LSTM model.

LSTM hyperparameters	Values
Batch size	1 / 250 / 13,742
Optimizer	Adam
Epoch	50 / 100 / 150 / 200
Loss	MSE / MAE / Hube loss
LSTM layers	2 / 3
LSTM cell per layer	256 / 128 / 64 / 32
Activation function	Linear / Tanh / Sigmoid / ReLU

2.6.3. Retraining

Once the soft sensor has been trained and evaluated with the legacy data (Fig. 4a), the subsequent step is to deploy the ML soft sensor in the pipeline (Fig. 4b). In this context, the time series database sends the data to the pre-processing module, which then transmits the processed data to the soft sensor inputs. This enables simulations of penicillin production as well as predictions for the following eight hours which can then be sent to a dashboard *(supplementary material)*.

Alternatively, the data can be sent to a retraining module where the soft sensor can be retrained to ensure its continued accuracy and efficacy over time. As new data becomes available, it can identify changes in patterns, trends, and behaviours that were not present in the original training data. Retraining is a relevant task used as a strategy to mitigate the impact of concept drift on a ML soft sensor. Where, concept drift refers to the statistical properties of the model target variable that change over time in an unexpected way (Widmer and Kubat, 1996). If concept drift occurs, then the pattern of past data may not be relevant to the new data and because the past data is the training set then this may lead the model to poor predictions and decision outcomes (Lu et al., 2019).

In a bioreactor, concept drift occurs mainly due to changes in microbial behaviour, variations in raw materials, equipment degradation, or modifications in operating conditions. For instance, changes in the composition of the fermentation medium or alterations in agitation and aeration rates can lead to shifts in the relationships between process variables. To monitor concept drift, it is recommended to assess the soft sensor model performance using a golden dataset. We defined a golden dataset as a reliable, well-annotated, high quality source of data for testing or benchmarking purposes. In this context, the soft sensor ML model is run on the golden dataset to predict penicillin production. These predictions are then compared to simulations generated by real-time data from the bioreactor using performance metrics (Fig. 4c).

Subsequently, performance metrics such as R^2, MAE, MSE, or RMSE should be monitored, to identify any significant deviations or trends indicating a drift. Finally, implementing thresholds for these performance metrics can automate the detection process, by acting as an alert to potential drift. For the *IndPenSim* case study, the batches containing deviations in the processes (91–100) were used to assess concept drift, which is shown in section 3.2.

3. MLOps pipeline in action

After understanding the most relevant components of MLOps for the construction of the soft sensor, the next logical questions are: does the soft sensor achieve good prediction quality under known scenarios? Is it necessary to retrain the soft sensor each time new data is received? How can drift be detected when there are deviations in the process? This section presents the evaluation of the soft sensor under both controlled experiments and experiments with deviations, as well as the analysis of model quality decay.

3.1. Classical evaluation of the sensor model with batch process deviations

The classical approach to evaluate a soft sensor model typically involves assessing its performance using a test set, consisting of unknown or unused data in the training set. To illustrate the conventional evaluation of soft sensors based on ML, in this section we have omitted IndPenSim batches 6, 13, 27, 33, 42 and 51 from the training phase. Subsequently, the omitted batches are used as a test set to calculate the statistical criteria for the conventional evaluation of the soft sensor (Table 3). The favourable results presented in Table 3 can be attributed

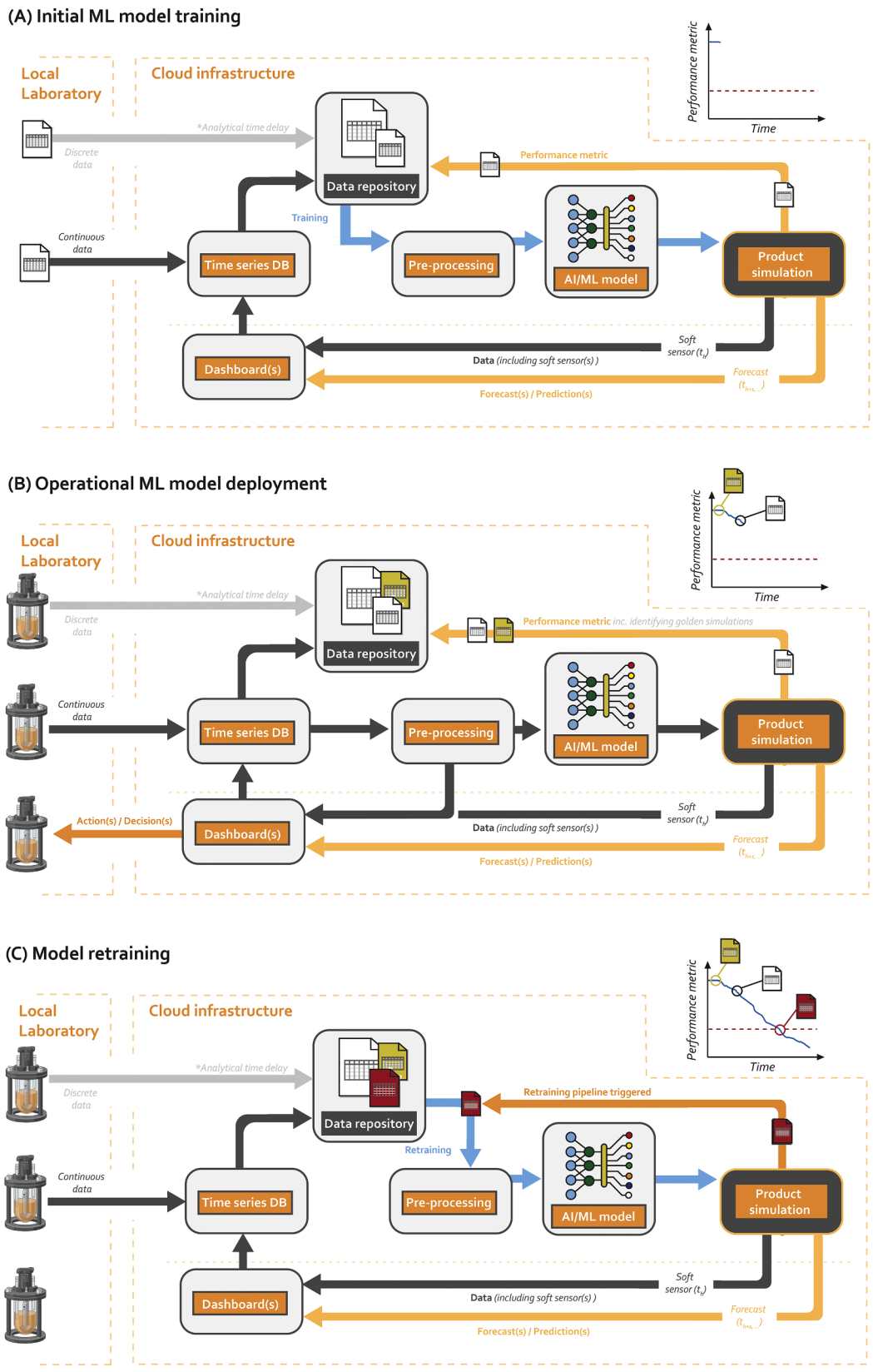

Fig. 4. Pipeline for monitoring the concept drift of the soft sensor based on golden dataset.

to fact that the data employed for training and testing are similar in nature (controlled by operator). Therefore, the ML model has been able to make accurate predictions by focusing on the recurrent patterns present within the batches. Whilst this suggests that the ML soft sensor can perform reliably under similar conditions it does not guarantee that the soft sensor will be robust under different, or more challenging, conditions (*e.g.*, faults or deviations in the experiments).

In order to demonstrate the concept drift in a dynamic domain as

Table 3
Results of the soft sensor evaluated with batches: 6, 13, 27, 33, 42 and 51. Statistical criteria used to estimate the performance of the LSTM model: R2, MAE, MSE, and RMSE.

Batch	R^2	MAE	MSE	RMSE
6	0.971	0.970	1.542	1.242
13	0.956	1.509	4.674	2.161
27	0.960	1.098	3.173	1.781
33	0.964	1.242	2.314	1.521
42	0.972	0.810	1.068	1.033
51	0.961	0.933	1.192	1.091

bioprocess, the soft sensor was trained on batches 1–60 which do not contain faults or deviations in the process and it was evaluated with 10 test datasets which correspond to IndPenSim batches (91–100) with process deviations. As illustrated in Table 4, three IndPenSim batches (94, 99 and 100) indicate a model decay in the performance of the soft sensor, with values of the coefficient of determination below 0.327 and elevated error metrics.

The error metrics represent the following: MAE is the average absolute difference between the predicted and actual penicillin concentration; MSE measures the average squared difference between the predicted and actual penicillin concentrations (penalising larger deviations more than smaller ones); and RMSE is the square root of MSE, representing the standard deviation of the errors between the predicted and actual penicillin concentration. MSE and RMSE calculate errors by squaring the difference between the predicted values and the actual values. This means that these metrics penalise larger error values more severely, according to the squared value of the difference.

The results for MAE for IndPenSim batches 94, 99 and 100 reveal that on average the soft sensor predictions are off around 2.39–~2.68 [g L^{-1}] from the real values. Whereas, RMSE for IndPenSim batches 94, 99 and 100 are considerably high in comparison to the maximum observed values for penicillin production of 12.926 [g L^{-1}], 13.709 [g L^{-1}] and 11.633 [g L^{-1}] respectively. This indicates that the mean error (standard deviation of prediction errors) is approximately equal to the maximum value observed within the range of the data.

On the other hand, the determination coefficient measures the proportion of total variation of the penicillin concentration measurements that is explained by the penicillin concentration simulations generated by the soft sensor model. In other words, R^2 represents the proportion of the variance in the observed outcomes that is explained by the predictions of the soft sensor, indicating its ability to replicate the observed variability. As mentioned above, the R^2 values are <0.327 for IndPenSim batches 94, 99 and 100 revealing that the performance of the soft sensor model has decayed. Graphically, Fig. 5 presents a visual representation of the penicillin concentration simulated by ML soft sensor and penicillin concentration of the IndPenSim batches 92, 94, 99 and 100. It can be observed that the IndPenSim batches 94, 99, 100 do not provide meaningful predictions, suggesting that the penicillin concentration measurements are poorly simulated in comparison to IndPenSim batch 92, which has also the highest R^2 value (0.971) in Table 4.

Table 4
Results of the soft sensor evaluated with process deviations batches (91–100). Statistical criteria used to estimate the performance of the LSTM model: R^2, MAE, MSE, and RMSE.

Batch	R^2	MAE	MSE	RMSE
91	0.737	0.976	1.264	1.598
92	0.971	1.147	1.435	2.058
93	0.956	1.333	2.273	5.166
94	0.327	2.504	3.222	10.384
95	0.861	1.024	1.342	1.800
96	0.922	1.898	2.700	7.289
97	0.783	2.757	3.870	14.979
98	0.925	2.398	3.145	9.890
99	0.303	2.681	3.351	11.229
100	0.270	2.397	3.036	9.219

3.2. Evaluation of the concept drift and retraining strategy

It is evident that the soft sensor model lacks the ability to learn when failures occur due to process deviations. The goal is to ensure that the soft sensor model remains accurate and relevant as the underlying data distribution changes. Unlike classical evaluation with a fixed test set, the evolving nature of the data necessitates an evaluation of the concept drift as variation may lead to a decline in model performance.

Monitoring concept drift is crucial for maintaining the performance of the LSTM soft sensor over time, especially with batches with faulty data when distributions are most likely to change. To illustrate the monitoring of concept drift in a soft sensor, we employed as golden datasets, batches 92 and 93, which achieved the best performance by the soft sensor. Remains of the batches presented in Table 4 were included to evaluate the concept drift of the LSTM soft sensor. Monitoring of the concept drift is based on the baseline metric score. We used the Population Stability Index (PSI) as metric score to quantify covariate shifts between two distributions (Taplin and Hunt, 2019). We are interested in detecting changes in the distribution of penicillin production in our ML model over time, as this may lead to a decline in model performance.

Prior to calculating the PSI for continuous variables, it is necessary to discretise the data. This is achieved by splitting the data into n bins, which are then used to represent it as a histogram. In this study, Doane's formula is employed to determine the number of bins (Doane, 1976).

The calculation of the PSI is given by the following equation:

$$PSI = \sum_{i=1}^{n} \left((r_i - m_i) \cdot \ln\left(\frac{m_i}{r_i}\right) \right)$$

Where r_i is the proportion of the reference data in the i^{th} bin:

$$r_i = \frac{number\ of\ reference\ data\ points\ in\ bin\ i}{number\ of\ reference\ data\ points}$$

And m_i the proportion of the monitored data in the i^{th} bin:

$$m_i = \frac{number\ of\ monitored\ data\ points\ in\ bin\ i}{number\ of\ reference\ data\ points}$$

The output of PSI is commonly interpreted as either indicating: no significant population change ($PSI < 0.1$); moderate change to the distribution ($0.1 \leq PSI < 0.25$); or suggests significant changes to the distribution ($PSI \geq 0.25$). It is important to note that these values are intended as general guidelines and should not be regarded as definitive benchmarks.

In Table 5 a comparison was conducted between the two distributions of penicillin concentration using the PSI between the golden dataset and the predicted penicillin concentrations by the LSTM model. The golden dataset (batches 92 and 93) represents the data (excluding the training set) on which the soft sensor performs well and serves as a baseline for comparing new, real-time data. In order to ensure that the baseline reflects a range of different, representative scenarios and is not overly reliant on a single dataset, two golden datasets were used. The monitored data refers to the "real-time data" from batches 91, 94, 95, 96, 97, 98, 99, and 100, collected while the ML soft sensor is deployed in production.

The PSI values presented in Table 5 indicate a considerable degree of concept drift between the majority of monitored batches and the golden datasets. This is evidenced by values that are significantly in excess of the threshold of 0.25, which represents changes in the fermentation process (*e.g.*, a deviation in operational parameters). Batch 96 is notable for exhibiting the lowest PSI values (1.943 and 0.234), indicating a relatively closer alignment with the golden dataset's distributions, particularly that of dataset 93. This indicates that batch 96 may represent a less divergent case, whereas other batches (*e.g.*, 91, 94, 95, 99 and

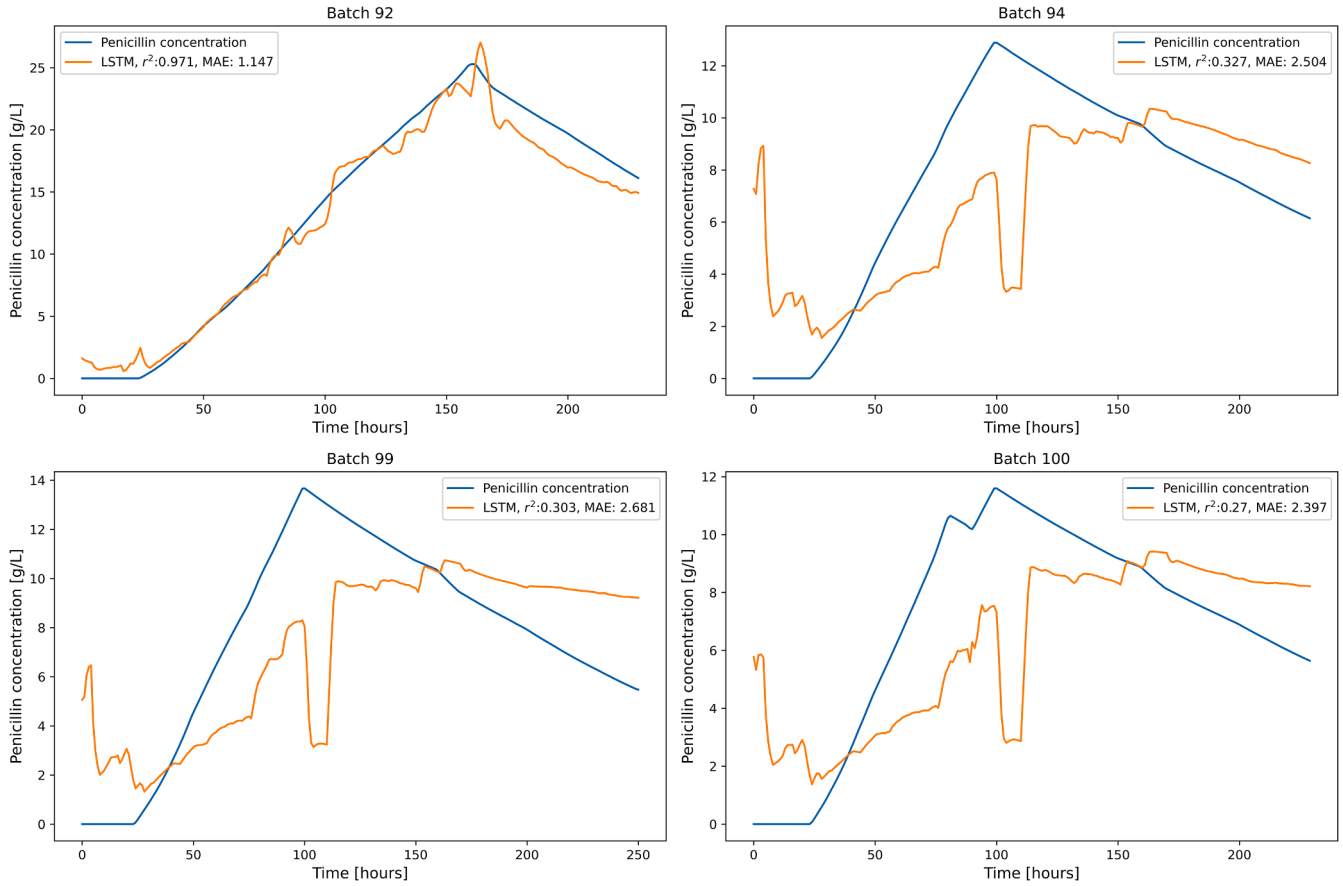

Fig. 5. Determination coefficient for IndPenSim batches 92, 94, 99, and 100. The IndPenSim batch 92 exhibited the highest determination coefficient, whereas the IndPenSim batches 94, 99, and 100 demonstrated a model decay in the soft sensor model.

Table 5
Population Stability Index (PSI) for monitored batches 91, 94, 95, 96, 97, 98, 99 and 100. Golden datasets 92 and 93. The simulated penicillin by LSTM is the variable monitored.

	Monitored batches							
Golden dataset	91	94	95	96	97	98	99	100
92	9.314	8.486	8.188	1.943	3.147	5.180	8.489	9.238
93	9.135	8.035	7.742	0.234	4.754	2.715	8.567	8.300

100) consistently demonstrate high PSI values, reflecting greater deviations. Consequently, these batches require detailed analysis to precisely identify the deviations in the fermentation process (section 3.3).

Upon detecting substantial alterations in the ML soft sensor simulations, indicated by PSI values exceeding 0.25, an alert is automatically triggered, notifying that concept drift has occurred. In this context, a retraining pipeline is scheduled once the fermentation batch has been completed. As the ML soft sensor is a recurrent neural network (LSTM), the model is not trained from scratch but is instead updated. During this update, the model's performance is adjusted by fine-tuning two key hyperparameters: epochs and batch size.

Setting the number of epochs ensures that the model has sufficient opportunities to learn from the new data, particularly in cases where deviations in the fermentation process may have led to changes in the relationships between input features and output predictions. On the other hand, the batch size determines the number of samples the model will process before updating the weights. The fine-tuning of the LSTM was performed semi-automatically, using a grid search of hyperparameter values and verified by an ML engineer.

The data pipeline approach for retraining, inspired by the work of (Dessaigne et al., 2024), was implemented with the objective of facilitating the deployment of data workflows. This approach employed Apache Airflow to orchestrate the various stages of the pipeline, including the connection to the time-series database, the data preprocessing module, and the training, evaluation, and retraining processes of the ML soft sensor. (Dessaigne et al., 2024) showcased the benefits of Airflow for coordinating complex pipelines, providing a ETL pipeline for addressing the legacy systems in fermentation process in real-time.

3.3. Analysis of the concept drift

To precisely identify the cause of the deviations in the fermentation process for the identified concept drift batches requires detailed analysis. IndPenSim batches 91 and 100 present disturbances in the variables sugar flow rate (F_s) and aeration both of which directly affect penicillin concentration (Fig. 6). For example, IndPenSim batch 91 exhibited faults in F_s at approximately 20 and 80 h into the process, indicating that the penicillin concentration will be inferior to that observed in other batches. Similar deviations were also observed in IndPenSim batches 95 and 97.

In contrast, IndPenSim batch 100 has faults in aeration occurring

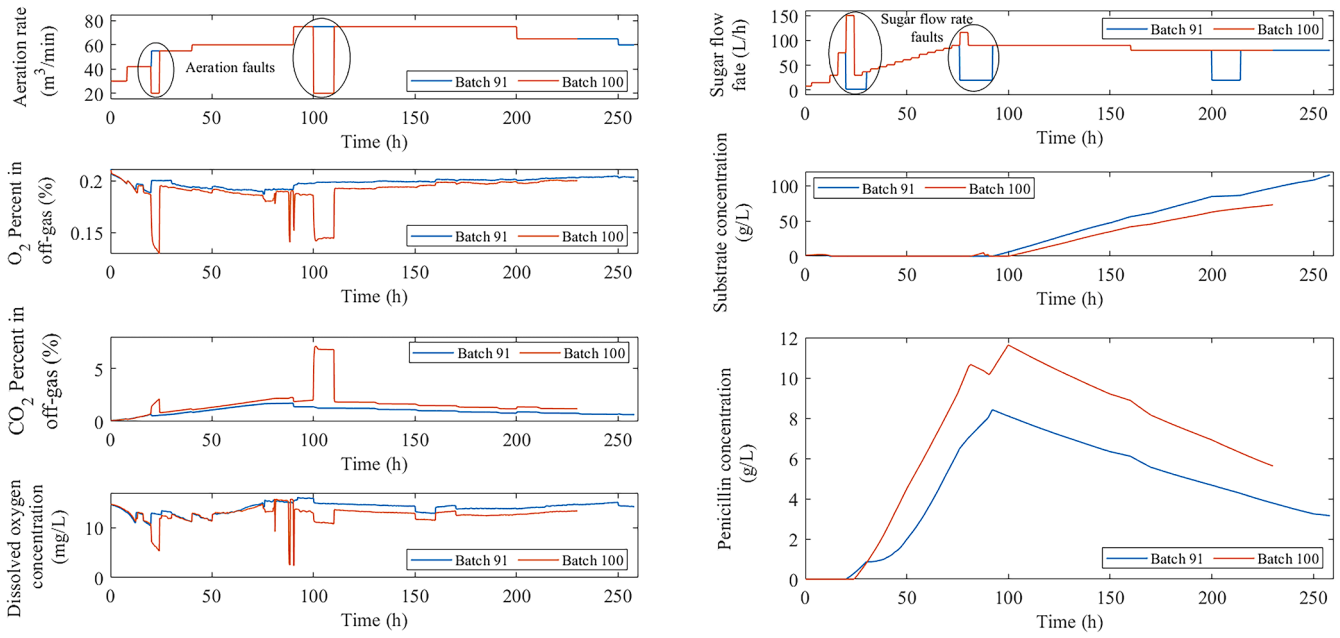

Fig. 6. Disturbance analysis. Variables with disturbances: aeration rate and sugar flow rate. Variables used in training: CO_{2out}, O_{2out}, and DO_2. Other variables: Substrate concentration. Predicted variables: Penicillin concentration. Details regarding the distributions associated with both the training and test sets for the soft sensor variables can be found in Figure A7–Figure A9 in the supplementary material.

around 20 and 100 h into the fermentation. Such a fault reduces the system's DO_2, leading to a decrease in O_{2out} and increase in CO_{2out}. A similar problem occurs in IndPenSim batches 94 and 99. The concentration of penicillin shows two marked peaks around 81 and 100 h for IndPenSim batch 100, which could be related to the disturbances in the substrate flow rate. Furthermore, substrate accumulation occurs in the system in both IndPenSim batches 91 and 100, which could be associated with substrate inhibition leading to a gradual decrease in the penicillin concentration is observed in both cases.

4. Conclusions and future works

In this paper, a comprehensive overview of the components necessary for a MLOps pipeline has been outlined alongside details of a proposed soft sensor. As an example, a proof of concept for MLOps is proposed to automate the end-to-end lifecycle of a soft sensor in industrial scale fed-batch fermentation, covering the development, deployment, maintenance and monitoring. As a use case, we utilised 100 batches of *Penicillium chrysogenum* in a 100,000 litre bioreactor (Goldrick et al., 2015).

The main contribution of this paper is that it provides a comprehensive overview of the components of an ML pipeline. It outlines the construction of a soft sensor to simulate penicillin production in real-time and is able to predict penicillin concentration over an 8 h forecast horizon. We have employed a LSTM learner for the development of the soft sensor, a popular recurrent neural network in recent years. The LSTM is equivalent to the nonlinear autoregressive model for time series forecasting problems (Zhang et al., 1998). In this context, the LSTM soft sensor incorporates predictor variables as time-lagged observations from the sensors (see Table 1: T, pH, DO_2, O_{2out}, CO_{2out}) and actuators (see Table 1: F_s, S, F_{oil}, NH_3_shots) of a bioreactor, excluding time-lagged observations of the target (penicillin concentration), a common practice in traditional time series problems. However, the utilisation of the LSTM serves to illustrate the construction of a soft sensor, as a key component in the creation of MLOps. The choice of learner can be adapted to align with the specific objectives of the soft sensor, whether that be a more transparent (*e.g.,* tree-based structure or rule-based representation) or a more black box learner alternative (*e.g.,* support vector machines, ensemble methods, transformers, etc.).

Furthermore, MLOps offers guidelines for the deployment, maintenance, and monitoring of the soft sensor in scenarios where data is generated in real-time with disturbances in the bioprocess (IndPenSim batches 91–100). Additionally, the dynamic nature of bioprocesses presents a significant challenge in maintaining ML soft sensors. Any changes in the data produced by sensors or actuators on which the soft sensor has been trained may result in a deterioration in performance. In this sense, the concept drift is the primary challenge in the soft sensor maintenance and monitoring. Concept drift might be attributed to changes such as degradation in the quality of materials of the system's equipment, seasonality, changing personal preferences and behaviours, or adversarial activities (Barros and Santos, 2019).

Based on the mentioned findings, we suggest as future research directions:

- Explore other approaches for concept drift detection based on statistical process control (SPC), window-based detectors, and ensemble learning. SPC monitors statistical properties over time to identify deviations, while window-based detectors compare statistical properties between fixed or adaptive time windows. Ensemble learning operate by combining the results of multiple diverse base learners. The overall performance is monitored by either considering the accuracy of all the ensemble members or the accuracy of each individual base learner (Bayram et al., 2022). Our objective is to develop an ensemble learning approach based on LSTM (H. Wang et al., 2021). This method consists of combining a sequence of LSTM weak learners to construct a strong learner. During the integration process, the weights of the weak learners will be adjusted based on their predicted losses. This approach will form the basis for further investigation and enhancement for detecting concept drift.
- The deployment of MLOps soft sensors marks the initial stage in the development of digital twins for bioprocesses, offering the potential for more effective monitoring, control and optimisation. Integrating data from soft sensors allows the creation of real-time virtual models of the bioprocess. The integration of digital twins (DT) facilitates the two-way flow of information between the physical and virtual worlds. We suggest that future work should adopt a DT approach,

whereby real-time sensors are coupled to the bioreactor to update the state of various parameters necessary for process or mechanistic modelling or machine learning applications. This will enable the current or future conditions to be derived or predicted. This allows for the testing of different control strategies in near real-time, enabling the steering of a process towards a specific objective (*e.g.,* increased yield, reduced cost, etc.). It is an approach gaining momentum in a number of disciplines where decision making is complex and information needs to be viewed in a holistic way (Tao and Qi, 2019; Metcalfe et al., 2023; "The Increasing Potential and Challenges of Digital Twins," 2024)

CRediT authorship contribution statement

Brett Metcalfe: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Juan Camilo Acosta-Pavas:** Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Carlos Eduardo Robles-Rodriguez:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **George K. Georgakilas:** Writing – original draft, Formal analysis, Conceptualization. **Theodore Dalamagas:** Writing – review & editing, Validation, Conceptualization. **Cesar Arturo Aceves-Lara:** Writing – review & editing, Validation, Conceptualization. **Fayza Daboussi:** Writing – review & editing, Validation, Conceptualization. **Jasper J Koehorst:** Writing – review & editing, Methodology, Conceptualization. **David Camilo Corrales:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

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.

Acknowledgements

This work was funded (or co-funded) by the European Union under the Horizon Europe project Bioindustry 4.0, grant n. 101094287.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2024.108991.

Data availability

Data will be made available on request.

References

Acosta-Pavas, J.C., Robles-Rodriguez, C.E., Griol, D., Daboussi, F., Aceves-Lara, C.A., Corrales, D.C., 2024. Soft sensors based on interpretable learners for industrial-scale fed-batch fermentation: learning from simulations. Comput. Chem. Eng. 187, 108736. https://doi.org/10.1016/j.compchemeng.2024.108736.

Barros, R.S.M.de, Santos, S.G.T.de C, 2019. An overview and comprehensive comparison of ensembles for concept drift. Inf Fusion 52, 213–244. https://doi.org/10.1016/j.inffus.2019.03.006.

Bayram, F., Ahmed, B.S., Kassler, A., 2022. From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl. Based. Syst. 245, 108632. https://doi.org/10.1016/j.knosys.2022.108632.

Birol, G., Ündey, C., Çinar, A., 2002. A modular simulation package for fed-batch fermentation: penicillin production. Comput. Chem. Eng. 26 (11), 1553–1565. https://doi.org/10.1016/S0098-1354(02)00127-8.

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470.

Cardillo, A.G., Castellanos, M.M., Desailly, B., Dessoy, S., Mariti, M., Portela, R.M.C., Scutella, B., von Stosch, M., Tomba, E., Varsakelis, C., 2021. Towards *in silico* process modeling for vaccines. Trends Biotechnol. 39 (11), 1120–1130. https://doi.org/10.1016/j.tibtech.2021.02.004.

Claßen, J., Aupert, F., Reardon, K.F., Solle, D., Scheper, T., 2017. Spectroscopic sensors for in-line bioprocess monitoring in research and pharmaceutical industrial application. Anal. Bioanal. Chem. 409 (3), 651–666. https://doi.org/10.1007/s00216-016-0068-x.

Clomburg, J.M., Crumbley, A.M., Gonzalez, R., 2017. Industrial biomanufacturing: the future of chemical production. Science (1979) 355 (6320), aag0804. https://doi.org/10.1126/science.aag0804.

Corrales, D.C., Corrales, J.C., Ledezma, A., 2018. How to address the data quality issues in regression models: a guided process for data cleaning. Symmetry. (Basel) 10 (4). https://doi.org/10.3390/sym10040099. Article 4.

Corrales, D.C., Ledezma, A., Corrales, J.C., 2020. A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks. Appl. Soft. Comput. 90, 106180. https://doi.org/10.1016/j.asoc.2020.106180.

Dessaigne, A., Briane, A., Daboussi, F., Cescut, J., Corrales, D.C., 2024. From legacy systems to data pipelines modernization in fermentation process. In: 34th European Symposium on Computer Aided Process Engineering /15th International Symposium on Process Systems Engineering, 53, pp. 3205–3210.

Doane, D.P., 1976. Aesthetic frequency classifications. Am. Stat. 30 (4), 181–183.

Elman, J.L., 1990. Finding structure in time. Cogn. Sci. 14 (2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E.

Erickson, L.E., 2009. Bioreactors. Encyclopedia of Microbiology, 3rd ed. Academic Press, pp. 206–211.

Fortuna, L., Graziani, S., Rizzo, A., 2007. Soft sensors in industrial applications. In: Xibilia, M.G. (Ed.), Soft Sensors for Monitoring and Control of Industrial Processes. Springer, pp. 1–13. https://doi.org/10.1007/978-1-84628-480-9_1.

Goldrick, S. (2019). *Data for: Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process.* 1. https://doi.org/10.17632/pdnjz7zz5x.1.

Goldrick, S., Duran-Villalobos, C.A., Jankauskas, K., Lovett, D., Farid, S.S., Lennox, B., 2019. Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. Comput. Chem. Eng. 130, 106471. https://doi.org/10.1016/j.compchemeng.2019.05.037.

Goldrick, S., Ştefan, A., Lovett, D., Montague, G., Lennox, B., 2015. The development of an industrial-scale fed-batch fermentation simulation. J. Biotechnol. 193, 70–82. https://doi.org/10.1016/j.jbiotec.2014.10.029.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Optimization for training deep models. Deep Learning. MIT press.

Harris, F.J., 1978. On the use of windows for harmonic analysis with the discrete fourier transform. In: Proceedings of the IEEE, 66, pp. 51–83. https://doi.org/10.1109/PROC.1978.10837. Proceedings of the IEEE.

Helleckes, L.M., Hemmerich, J., Wiechert, W., Lieres, E., von, Grünberger, A., 2023. Machine learning in bioprocess development: from promise to practice. Trends Biotechnol. 41 (6), 817–835. https://doi.org/10.1016/j.tibtech.2022.10.010.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735. Neural Computation.

Huang, J., Shimizu, H., Shioya, S., 2002. Data preprocessing and output evaluation of an autoassociative neural network model for online fault detection in virginiamycin production. J. Biosci. Bioeng. 94 (1), 70–77. https://doi.org/10.1016/S1389-1723(02)80119-0.

Ji, C., Ma, F., Wang, J., Sun, W., 2023. Profitability related industrial-scale batch processes monitoring via deep learning based soft sensor development. Comput. Chem. Eng. 170, 108125. https://doi.org/10.1016/j.compchemeng.2022.108125.

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven Soft Sensors in the process industry. Comput. Chem. Eng. 33 (4), 795–814. https://doi.org/10.1016/j.compchemeng.2008.12.012.

Kaneko, H., Arakawa, M., Funatsu, K., 2009. Development of a new soft sensor method using independent component analysis and partial least squares. Aiche Journal 55, 87–98.

Lawrence, N.P., Damarla, S.K., Kim, J.W., Tulsyan, A., Amjad, F., Wang, K., Chachuat, B., Lee, J.M., Huang, B., Bhushan Gopaluni, R., 2024. Machine learning for industrial sensing and control: a survey and practical perspective. Control Eng Pract 145, 105841. https://doi.org/10.1016/j.conengprac.2024.105841.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G., 2019. Learning under concept drift: a review. IEEe Trans. Knowl. Data Eng. 31 (12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857. IEEE Transactions on Knowledge and Data Engineering.

Mandis, M., Baratti, R., Chebeir, J., Tronci, S., Romagnoli, J.A., 2024. Exploring nontraditional LSTM architectures for modeling demethanizer column operations. Comput. Chem. Eng. 183, 108591. https://doi.org/10.1016/j.compchemeng.2024.108591.

Meadows, A.L., Hawkins, K.M., Tsegaye, Y., Antipov, E., Kim, Y., Raetz, L., Dahl, R.H., Tai, A., Mahatdejkul-Meadows, T., Xu, L., Zhao, L., Dasika, M.S., Murarka, A., Lenihan, J., Eng, D., Leng, J.S., Liu, C.-L., Wenger, J.W., Jiang, H., Tsong, A.E., 2016. Rewriting yeast central carbon metabolism for industrial isoprenoid production. Nature 537 (7622), 694–697. https://doi.org/10.1038/nature19769.

Metcalfe, B., Boshuizen, H.C., Bulens, J., Koehorst, J.J., 2023. Digital Twin Maturity levels: A theoretical Framework For Defining Capabilities and Goals in the Life and Environmental Sciences, 12. F1000Research, p. 961. https://doi.org/10.12688/f1000research.137262.1.

Meyer, V., Basenko, E.Y., Benz, J.P., Braus, G.H., Caddick, M.X., Csukai, M., de Vries, R.P., Endy, D., Frisvad, J.C., Gunde-Cimerman, N., Haarmann, T., Hadar, Y., Hansen, K., Johnson, R.I., Keller, N.P., Kraševec, N., Mortensen, U.H., Perez, R., Ram, A.F.J., Wösten, H.A.B., 2020. Growing a circular economy with fungal biotechnology: a white paper. Fungal. Biol. Biotechnol. 7 (1), 5. https://doi.org/10.1186/s40694-020-00095-z.

Naseri, G., 2023. A roadmap to establish a comprehensive platform for sustainable manufacturing of natural products in yeast. Nat. Commun. 14 (1), 1916. https://doi.org/10.1038/s41467-023-37627-1.

Nielsen, J., Larsson, C., van Maris, A., Pronk, J., 2013. Metabolic engineering of yeast for production of fuels and chemicals. Curr. Opin. Biotechnol. 24 (3), 398–404. https://doi.org/10.1016/j.copbio.2013.03.023.

Paul, G.C., Thomas, C.R., 1996. A structured model for hyphal differentiation and penicillin production using penicillium chrysogenum. Biotechnol. Bioeng. 51 (5), 558–572. https://doi.org/10.1002/(SICI)1097-0290(19960905)51:5<558::AID-BIT8>3.0.CO;2-B.

Qiu, K., Wang, J., Zhou, X., Guo, Y., Wang, R., 2020. Soft sensor framework based on semisupervised just-in-time relevance vector regression for multiphase batch processes with unlabeled data. Ind. Eng. Chem. Res. 59 (44), 19633–19642. https://doi.org/10.1021/acs.iecr.0c03806.

Qiu, K., Wang, J., Zhou, X., Wang, R., Guo, Y., 2022. Soft sensor based on localized semi-supervised relevance vector machine for penicillin fermentation process with asymmetric data. Measurement 202, 111823. https://doi.org/10.1016/j.measurement.2022.111823.

Quinlan, J.R., 1992. Learning with continuous classes 92, 343–348.

Saisana, M., 2014. Standard Scores. In: Michalos, A.C. (Ed.), Encyclopedia of Quality of Life and Well-Being Research. Springer, Netherlands, pp. 6321–6322. https://doi.org/10.1007/978-94-007-0753-5_2852.

Shayegh, S., Reissl, S., Roshan, E., Calcaterra, M., 2023. An assessment of different transition pathways to a green global economy. Commun. Earth. Environ. 4 (1), 1–12. https://doi.org/10.1038/s43247-023-01109-5.

Siegl, M., Brunner, V., Geier, D., Becker, T., 2022. Ensemble-based adaptive soft sensor for fault-tolerant biomass monitoring. Eng. Life Sci. 22 (3–4), 229–241. https://doi.org/10.1002/elsc.202100091.

Siegl, M., Kämpf, M., Geier, D., Andreeßen, B., Max, S., Zavrel, M., Becker, T., 2023. Generalizability of soft sensors for bioprocesses through similarity analysis and phase-dependent recalibration. Sensors 23 (4), 4. https://doi.org/10.3390/s23042178.

Smiatek, J., Jung, A., Bluhmki, E., 2020. Towards a digital bioprocess replica: computational approaches in biopharmaceutical development and manufacturing. Trends Biotechnol. 38 (10), 1141–1153. https://doi.org/10.1016/j.tibtech.2020.05.008.

Stulp, F., Sigaud, O., 2015. Many regression algorithms, one unified model: a review. Neural Networks 69, 60–79. https://doi.org/10.1016/j.neunet.2015.05.005.

Tao, F., Qi, Q., 2019. Make more digital twins. Nature 573 (7775), 490–491. https://doi.org/10.1038/d41586-019-02849-1.

Taplin, R., Hunt, C., 2019. The population accuracy index: a new measure of population stability for model monitoring. Risks 7 (2), 2. https://doi.org/10.3390/risks7020053.

The increasing potential and challenges of digital twins, 2024. Nat. Comput. Sci. 4 (3), 145–146. https://doi.org/10.1038/s43588-024-00617-4.

Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1 (Jun), 211–244.

Treveil, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M., Heidmann, L., 2020. Introducing MLOps. O'Reilly Media.

Vojinović, V., Cabral, J.M.S., Fonseca, L.P., 2006. Real-time bioprocess monitoring: part I: in situ sensors. Sens. Actuators B. Chem. 114 (2), 1083–1091. https://doi.org/10.1016/j.snb.2005.07.059.

Wang, B., Nie, Y., Zhang, L., Song, Y., Zhu, Q., 2023. An soft-sensor method for the biochemical reaction process based on LSTM and transfer learning. Alex. Eng. J. 81, 170–177. https://doi.org/10.1016/j.aej.2023.09.007.

Wang, H., Li, M., Yue, X., 2021. IncLSTM: incremental ensemble LSTM model towards time series data. Comput. Electr. Eng. 92, 107156. https://doi.org/10.1016/j.compeleceng.2021.107156.

Wang, Y., Witten, I.H., 1997. Inducing model trees for continuous classes 9 (1), 128–137.

Widmer, G., Kubat, M., 1996. Learning in the presence of concept drift and hidden contexts. Mach. Learn. 23 (1), 69–101. https://doi.org/10.1007/BF00116900.

Yan, W., Tang, D., Lin, Y., 2017. a data-driven soft sensor modeling method based on deep learning and its application. IEEE Trans. Industr. Electron. 64 (5), 4237–4245. https://doi.org/10.1109/TIE.2016.2622668. IEEE Transactions on Industrial Electronics.

Zhang, G., Eddy Patuwo, B., Hu, Y., 1998. Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. 14 (1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7.

Zhang, J., Hansen, L.G., Gudich, O., Viehrig, K., Lassen, L.M.M., Schrübbers, L., Adhikari, K.B., Rubaszka, P., Carrasquer-Alvarez, E., Chen, L., D'Ambrosio, V., Lehka, B., Haidar, A.K., Nallapareddy, S., Giannakou, K., Laloux, M., Arsovska, D., Jørgensen, M.A.K., Chan, L.J.G., Keasling, J.D., 2022. A microbial supply chain for production of the anti-cancer drug vinblastine. Nature 609 (7926), 341–347. https://doi.org/10.1038/s41586-022-05157-3.