|  |  |
| --- | --- |
| cetlogo ***CHEMICAL ENGINEERING TRANSACTIONS***  ***VOL. xxx, 2025*** | A publication of  aidiclogo_grande |
| The Italian Association  of Chemical Engineering  Online at www.cetjournal.it |
| Guest Editors: David Bogle, Flavio Manenti, Piero Salatino  Copyright © 2025, AIDIC Servizi S.r.l. **ISBN** 979-12-81206-21-2; **ISSN** 2283-9216 | |

Soft-Sensors Strategy for Estimating the Cell Biomass and the Nanobody-Based Scorpion Antivenom Production: from ML Algorithms Selection to Scale-Down Bioprocess Application

Juan C. Acosta-Pavasa, David Camilo Corralesb, Susana M. Alonso Villelaa, Georgios Georgakilasc, Balkiss Bouhaouala-Zahard, Julien Cescutb, Luc Fillaudeaua, César Arturo Aceves-Laraa,\*

aTBI, Université de Toulouse, CNRS UMR5504, INRAE UMR792, INSA, Toulouse, France

bAthena Research Center, Marousi, Greece

cLaboratoire des Biomolécules, Venins et Applications Théranostiques (LBVAT), Institut Pasteur de Tunis, 13 Place Pasteur, BP-74, 1002 Le Belvédère, Tunis, Tunisia

d INRAE, UMS (1337) TWB, 135 Avenue de Rangueil, 31077 Toulouse, France

Corresponding author: [aceves@insa-toulouse.fr](mailto:aceves@insa-toulouse.fr)

The production of recombinant nanobodies (protein) in *E. coli* exhibits promising results for its application in serotherapy. However, the monitoring, control, and optimization of this bioprocess present limitations due to the measurements of critical variables. This study explores the potential of machine learning (ML) tools, such as soft sensors, to estimate offline variables such as protein and biomass concentration. CART Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Radial Basis Function (RBF) Support Vector Machine (SVM) learners were used to train and test the soft sensors. Eleven offline variables were used as features. For the protein soft sensor, KNN and RBF SVM exhibit the best performance with a training and test results , , , and . In the biomass soft sensor, the training of all the learners showed good results , , , and .The soft sensors were evaluated in a scale-down process to estimate the biomass using only four variables. The results show the potential to use soft sensors for monitoring variables that are not possible to measure in real time.

* 1. Introduction

Scorpion sting envenoming (scorpionism) is a severe public health issue in developing countries in North Africa, the Middle East, and Central and South America (Olivares-Hernández et al., 2024). The production of recombinant nanobodies in *E. coli* exhibits promising results for its application in serotherapy, especially for its implementation as a treatment for scorpion stings (Alonso Villela et al., 2021). However, the monitoring, control, and optimization of this bioprocess present limitations due to the measurements of critical variables. Soft Sensors have emerged as an alternative solution to this problem (Brunner et al., 2021).

Gundinger et al. (2022) developed a soft sensor for controlling the feeding strategy during the fed-batch cultivations for the production of a recombinant model antigen-binding fragment (Fab) in the periplasm of *E. coli* via phosphate (PO4)-sensitive phoA expression system (pAT). The soft sensor was trained to predict biomass concentration using four features: CO2 evolution rate, O2 uptake rate (measured by Off-gas analysis), the substrate feed rate, and the reactor biomass. Medl et al. (2023) developed a feed-forward neural network for real-time optical density estimation using volume-specific cumulative ammonia and volume-specific carbon feed additions as features. This soft sensor accurately predicted the optical density (>95%).

Although the works presented are promising, they primarily focus on biomass quantification, neglecting a deeper protein concentration analysis. Additionally, a more in-depth exploration of the explainability of different models could significantly enhance these studies. Methodologies such as Shapley Additive Explanations (SHAP) could be applied to gain insights into the model's decision-making process.

This work aims at exploring the potential of ML soft sensors to estimate protein and biomass concentration using different learners such as CART Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Radial Basis Function (RBF) Support Vector Machine (SVM). The contributions of this work are summarized as follows:

* Different machine learning (ML) algorithms were applied to build soft sensors to monitor the protein and biomass concentration
* The experimental protein data were augmented by a dynamic hybrid model (DHM) for nanobody-based antivenom (protein) production and symmetrically aligned with online measurements
* A feature importance using SHAP values was applied to study the effect of the most influential features in the training of soft sensors
* The use of biomass soft sensors was explored in an independent test dataset to check their accuracy for online monitoring
  1. Workflow of Soft Sensors Development
     1. Experimental Data

Seven experiments were conducted for protein production using two recombinant strains of E. coli WK6: CH10-12 and NbF 12-10. The working volume was (, ). Different induction temperatures were applied. Table 1 summarizes the main conditions reported by Alonso Villela et al. (2021). Protein content was monitored over time for each experiment, with a total between 15 and 28 experimental measurements. The bioreactor was equipped with a Multifunctional Controller Software BioPAT® MFCS (Sartorius) software, capable of measuring every 5 seconds online variables such as pH, pO2, pressure, etc., and a gas analyzer (INNOVA 1313, Lumasense Technologies) measuring every 5 minutes. The induction was promoted in each experiment by adding of isopropyl- thiogalactopyranoside (IPTG) to start protein production.

Table 1: Protein production experimental conditions (Alonso Villela et al., 2021).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Experiment | *E. coli* Strain | Initial glucose concentration  () | Glucose feed concentration () | Batch and fed-batch phase temperature (°C) | Induction phase conditions duration (h)-T (°C) |
| 1 | CH10-12 | 10 | 298 | 37 | 12h-28°C |
| 2 | 10 | 298 | 10h-30°C |
| 3 | 10 | 326 | 6h-33°C |
| 4 | 10 | 326 | 6h-37°C |
| 5 | 12.5 | 305 | 38h-29°C |
| 6 | 7.5 | 305 | 35h-32°C |
| 7 | NbF12-10 | 10 | 305 | 37h-29°C |

* + 1. Online and offline Data Integration and Treatment

The online and offline measurements are not in a one-to-one correspondence, e.g., the quantity of pH measurements per process time differs from the protein and biomass measurements performed offline, which is referred to as asymmetric data. To build the protein soft sensor, the experimental protein and biomass data were augmented by a DHM for protein production proposed by Corrales et al. (2024) and symmetrically aligned with the online measurements. The DHM accounts for biomass growth, glucose concentration, protein concentration, and changes in volume, considering different induction temperatures (Corrales et al., 2024). In this study, 2001 data points (instances) per experiment were simulated. Each instance for protein and biomass concentration was symmetrically aligned with corresponding values from the offline dataset (), using the process time as a reference. If the time associated with the protein concentration did not exactly match the offline variable's time, the nearest preceding time was selected. Subsequently, the datasets were filtered using a moving average filter to reduce the noise that could influence the training of the ML algorithms.

* + 1. Training and Test of Protein and Biomass Soft Sensors

The construction of protein and biomass soft sensors was divided into the following steps,

1. A split 80/20 was performed over both datasets: 80% was used to train the learners and 20% to test
2. The protein and biomassdatasets were standardized using a Z-score
3. A (Pearson) correlation analysis was conducted to determine linear relationships between the studied features in the training set
4. Training: Four ML algorithms were trained to quantify the protein and biomass concentration: CART DT, RF, KNN, and RBF SVM. A 5-fold cross-validation with three repetitions was proposed to validate the learning process of the ML models. Protein/biomass dataset size (1120511)
5. Test: It was performed a test to evaluate the soft sensors performance. Protein/biomass dataset size (280211)

The performance of protein and biomass soft sensors were analyzed using the Mean Absolute Error , [Mean Squared Error](https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/) , [Root Mean Squared Error](https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/) , and Coefficient of Determination .

* + 1. Feature Importance (SHAP Values)

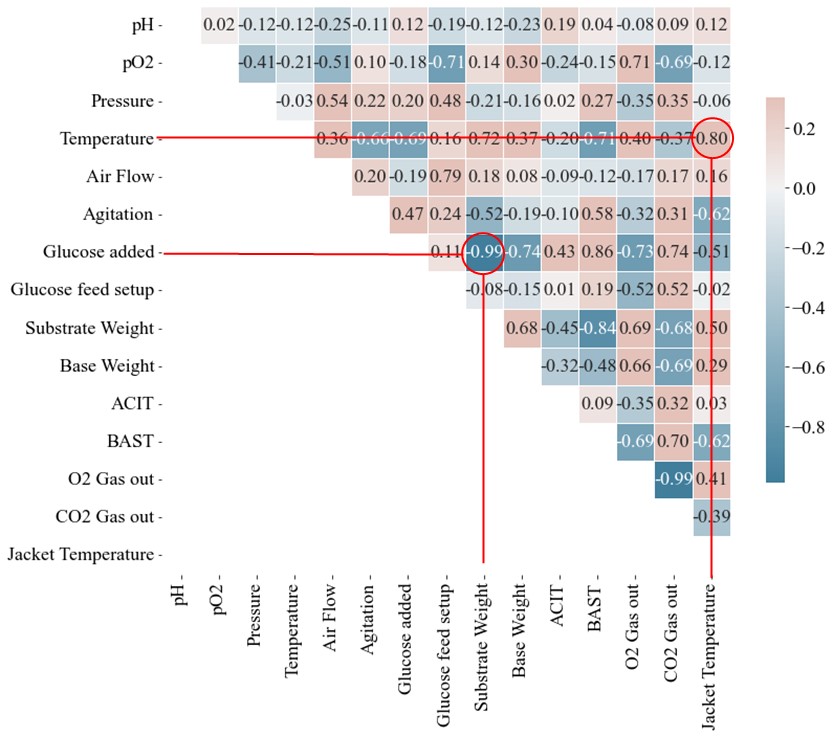
Shapley Additive exPlanations (SHAP) is a technique based on cooperative game theory used as a *post hoc* (explainability) method to explain how an ML model makes individual predictions. SHAP is a model-agnostic technique that can be applied to any ML algorithm (Molnar, 2022). This increases the interpretability of ML models by decomposing the prediction into a sum of contributions from each of the model's input variables (features). The SHAP values provide local and global interpretation. This work is oriented on the global interpretation by focusing on the impact and importance of the features in the model predictions (Molnar, 2022). The SHAP values explain the model’s performance (predictions) and must be interpreted in the context of the data used to train the learners. SHAP values were employed to study the most influential online variables in training the protein and biomass soft sensors using CART DT, RF, KNN, and RBF SVM as learners. To reduce computational costs, instances were grouped into clusters using a weighted k-means (k=10) approach to compute the SHAP values for the protein and biomass soft sensors. Instead of calculating SHAP values for each instance, the instances were clustered, and SHAP values were computed for the centroids of these clusters, with each centroid weighted by the number of data points it represented.

* + 1. Soft Sensors Scale-Down Test

The biomass soft sensors were applied in a scale-down strategy; the biomass concentrations were estimated using a robotic bioprocessing platform that integrates 24 bioreactors with a working volume of operating at with *E. coli* CH10-12 for the batch, fed-batch ( and the induction phases. Two of these bioreactors were used to evaluated the soft sensors: independent test set 1 and independent test set 2.

* 1. Results and Discussion
     1. Training and Test of Biomass and Protein Soft Sensors

All simulations were run using a 13th Gen Intel® Core(TM) i7-13800H 2.50 GHz, 32 GB RAM computer. The s*cikit-learn: Machine Learning* Python module was used to train and test the soft sensors using CART DT, RF, KNN, and RBF SVM as learners. The first step involved investigating the correlation between the 15 features presented in Figure 1 by calculating the Pearson correlation coefficient for each pair of features.



*Figure 1:* Pairwise *offline variables (*feature) correlations. -1 perfect negative correlation; 0 no correlation; +1 perfect positive correlation.

The objective of applying the Pearson correlation coefficient between pairs of variables is to reduce them in the training if they are highly correlated. In this work, a high correlation is assumed if a pairwise (, ) exhibits , which would suggest that one of them should be removed from the training dataset. This selection can be justified based on the expert's experience and the importance assigned to the variable from an experimental point of view. Figure 1 shows that temperature correlates with jacket temperature (0.80); O2, Gas out correlates with CO2, Gas out (-0.99); Glucose added correlates with substrate weight (-0.99) and BAST (0.86); Substrate weight also presents a strong negative correlation with BAST (-0.84). Substrate weight can be easily measured in a bioprocess, while glucose added requires some extra calculations. Temperature and Jacket temperature are variables that from experimental point of view reveals similar information. A similar conclusion could be obtained from the pairwise O2, Gas out; CO2, Gas out and Glucose added; substrate weight. In the experiments developed, the addition of glucose results in the reduction of the pH in the process. The pH of the feed medium is ~2-3 (Alonso Villela et al., 2021), which means that the variable BAST will increase. Based on this analysis and the Pearson correlation results, the following variables are removed: Jacket temperature, CO2, Gas out, BAST, and Glucose added. Table 2 displays the statistical results in training and test of different ML algorithms using eleven features. For the protein soft sensor, KNN and RBF SVM exhibit the best performance with a training and test results , , , and . In the biomass soft sensor, the training of all the learners showed good results , , , and .

Table 2: Statistical analysis in the training and test of different ML algorithms using 11 variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Soft sensor** | **Metrics/Model** | **CART DT** | **RF** | **KNN** | **RBF SVM** |
|  |  | Training | | | |
| *Protein* |  | 0.980 | 0.987 | 0.999 | 0.997 |
|  | 0.020 | 0.013 | 1.009×10-6 | 3.482×10-3 |
|  | 0.140 | 0.113 | 1.004×10-3 | 0.059 |
|  | 0.082 | 0.568 | 6.378×10-4 | 0.052 |
|  | Test | | | |
|  | 0.980 | 0.986 | 0.999 | 0.996 |
|  | 0.022 | 0.013 | 4.958×10-6 | 3.578×10-3 |
|  | 0.618 | 0.115 | 2.226×10-3 | 0.060 |
|  | 0.087 | 0.071 | 5.393×10-4 | 0.052 |
| *Biomass* |  | Training | | | |
|  | 0.992 | 0.994 | 0.999 | 0.995 |
|  | 7.569×10-3 | 6.253×10-3 | 7.001×10-5 | 4.628×10-3 |
|  | 0.087 | 0.079 | 8.368×10-3 | 0.068 |
|  | 0.059 | 0.054 | 2.178×10-3 | 0.062 |
|  | Test | | | |
|  | 0.992 | 0.994 | 0.999 | 0.995 |
|  | 8.356×10-3 | 6.113×10-3 | 1.081×10-4 | 4.630×10-3 |
|  | 0.091 | 0.078 | 0.010 | 0.068 |
|  | 0.061 | 0.054 | 2.858×10-3 | 0.062 |

Figure 2 illustrates the results of predicting the protein and biomass concentrations in the training and test datasets of experiment 7. In Figure 2a, the protein soft sensor used KNN as a learner; meanwhile, in Figure 2b, the biomass soft sensors used RBF SVM. The example illustrates that the soft sensor satisfactory estimates the protein and biomass concentration of two strains of *E. coli* submitted at different operational conditions.

|  |  |
| --- | --- |
| (a) KNN protein soft sensor prediction | (b) RBF SVM biomass soft sensor prediction |

*Figure 2: Protein (KNN learner) and biomass (RBF SVM learner) soft sensors predictions over experiment 7. Strain: E. coli* CH10-12; induction temperature: 29°C.

* + 1. SHAP Values

Figure 3 presents the Beeswarm plot, which permits global interpretability of the SHAP values, i.e., how features affect the prediction and their importance (Molnar, 2022). Figures 3a and 3b show the influence of the eleven features in the protein and biomass soft. The features are ranked by Mean absolute SHAP values, and the colors reveal how an increase (red) or decrease (blue) of the feature affects the predictions, e.g., higher values of substrate weight have negative SHAP values, which indicates that lower protein predictions are associated with higher substrate weights.

|  |  |
| --- | --- |
| (a) SHAP values protein soft sensor | (b) SHAP values biomass soft sensor |

*Figure 3: SHAP values (*Beeswarm plot) for *protein and biomass soft sensors. The SHAP values were computed for the four ML algorithms studied.*

* + 1. Biomass Scale-Down Process

The aim of this work was to evaluate the soft sensors applied to scale down process compared to the initial experiments . Scale-down process have lower quantity of online measurements. Based on the SHAP values results, it was selected four online variables to retrain the biomass soft sensors: Temperature, pO2, agitation, and pH. The biomass training dataset was resized from (1120511) to (112054). This allows to use the retrained biomass soft sensor and evaluate its performance using independent test datasets 1 and 2. Table 3 displays the statistical results in training different ML algorithms using four features.

Table 3: Statistical analysis in the training and test of different ML algorithms using 4 variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Soft sensor** | **Metrics/Model** | **CART DT** | **RF** | **KNN** | **RBF SVM** |
| *Biomass* |  | Training | | | |
|  | 0.969 | 0.972 | 0.993 | 0.971 |
|  | 0.031 | 0.028 | 6.920×10-3 | 0.029 |
|  | 0.175 | 0.167 | 0.083 | 0.169 |
|  | 0.105 | 0.102 | 0.035 | 0.115 |

The statistical analysis reveals that the retraining with four features caused a slight decrease in all the metrics compared with the training using eleven variables. Nevertheless, the training results with less variable shows a good performance of the process at . It can be concluded that it is feasible to use four variables to train biomass soft sensors. Figure 4a displays the results of two independent experiments (independent test datasets 1 and 2) developed at a lower scale with the independent test set.

|  |  |
| --- | --- |
| (a) Biomass soft sensor predictions | (b) Density plots |

*Figure 4: (a) B*iomass concentration estimation over independent test sets 1 and 2 using different soft sensors. (b) density plots for online variables used in the scale-down test.

In the independent test set 1, most of the soft sensors presents , where DT and RBF SVM exhibit the best performance with a . In the independent test set 2, only the DT present acceptable results with . To understand better the results, Figure 4b shows the density plots of the training datasets and the independent test dataset. Both datasets have differences in the online variables and the biomass predicted. In the training set, the instances for the feature agitation are mainly concentrated in values of 500, 1000, and 1500 . Meanwhile, in the independent test set, the instances for the same variable are mainly concentrated at 600 and 1250 . This is similar to the other variables, temperature, pO2, and pH. This fact is mainly related to changes in the operational conditions, such as feed concentration or induction temperature between scales.

* 1. Conclusions

ML soft sensors constitute a promising alternative in the field of bioprocesses. Their potential for estimating variables such as protein and biomass concentrations from commonly measured online variables makes them an attractive tool for rapid decision-making and real-time adjustments to estimate critical parameters in bioprocess. Initial studies have shown interesting results in applying soft sensors as scale-down tools using learners such as DT, RF, KNN, or RBF SVM. Future work will explore the application of soft sensors for monitoring protein and biomass concentrations on an industrial scale.

Acknowledgments

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

References

Alonso Villela, S. M., Ghezal-Kraïem, H., Bouhaouala-Zahar, B., Bideaux, C., Aceves Lara, C. A., & Fillaudeau, L. (2021). Effect of temperature on the production of a recombinant antivenom in fed-batch mode. *Applied Microbiology and Biotechnology*, *105*(3), 1017‑1030.

Brunner, V., Siegl, M., Geier, D., & Becker, T. (2021). Challenges in the Development of Soft Sensors for Bioprocesses: A Critical Review. Frontiers in Bioengineering and Biotechnology, 9, 722202.

Corrales, D. C., Villela, S. M. A., Bouhaouala-Zahar, B., Cescut, J., Daboussi, F., O’donohue, M., Fillaudeau, L., & Aceves-Lara, C. A. (2024). Dynamic Hybrid Model for Nanobody-based Antivenom Production (scorpion antivenon) with E. coli CH10-12 and E. coli NbF12-10. In Computer Aided Chemical Engineering (Vol. 53, p. 145‑150). Elsevier.

Gundinger, T., Kittler, S., Kubicek, S., Kopp, J., & Spadiut, O. (2022). Recombinant Protein Production in E. coli Using the phoA Expression System. Fermentation, 8(4), 181.

Medl, M., Rajamanickam, V., Striedner, G., & Newton, J. (2023). Development and Validation of an Artificial Neural-Network-Based Optical Density Soft Sensor for a High-Throughput Fermentation System. Processes, 11(1), 297.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Munich, Germany. Molnar.

Olivares-Hernández, R., Riaño-Umbarila, L., Becerril, B., Alagón, A., & Vázquez-López, H. (2024). Pharmacokinetic evaluation of a single chain antibody fragment against scorpion toxins in sheep. *Toxicon*, *242*, 107691.

Ponce‐Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis : Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11), e70056.