**LSTM-based soft sensor for the prediction of microalgae growth**

Tehreem Syed,a\* Shyam Kalliadan,b,c Jonathan Mädler,d Kris Laukens,c Luc Roef*,*b Leon Urbasd

*aFaculty of Electrical and Computer Engineering, Technische Universität Dresden, 01069 Dresden, Germany*

*bProviron Holding NV, 2620 Hemiksem, Belgium*

*cFaculty Department of Computer Science, University of Antwerp, 2020 Antwerpen, Belgium*

*dProcess Systems Engineering Group, Technische Universität Dresden, 01069 Dresden, Germany*

*tehreem.syed@tu-dresden.de*

Abstract

In biotechnological processes, biological process parameters such as biomass concentration, nutrient concentration, chlorophyll content, and product quality can be challenging to measure online. Typically, these parameters are measured in laboratory settings employing offline sample analyzers. Yet, offline measurements cannot be used as quick feedback signals for process control because of the significant delay between sampling and result generation. Moreover, though generally very accurate, they are equally expensive and often come with high maintenance costs. Therefore, soft sensors are widely utilized to address this problem, allowing reliable online estimations of these essential biological process parameters. Deep learning-based soft sensors are prevalent these days due to higher prediction performance. This work describes the use of Long Short-Term Memory (LSTM), a deep-learning architecture specifically designed for time series data, to predict microalgal growth. LSTM has the advantage of predicting future values based on history. The LSTM-based soft sensor developed for predicting microalgae biomass shows higher prediction performance than the support vector regression (SVR) based soft sensor. An LSTM was trained on the indoor cultivation data of *Nannochloropsis* cultivated in a vertical flat panel photobioreactor. The dataset consists of 28,741 samples, with 20,280 used for training and 8461 used for evaluation. Our LSTM-based soft sensor performed better than the SVR-based sensor and achieved an R2 score of 0.91, which was higher than the R2 score of 0.781 achieved using the SVR-based soft sensor.

**Keywords:** Microalgae cultivation, Soft sensor, Support vector regression, Long short-term memory

* 1. Introduction

Process parameters in microalgal cultivations that are often challenging to measure online are evaluated using offline laboratory analyzers. Although offline laboratory analyzers can provide accurate measurements, the sampling cycle is often very long, and it usually leads to small sampling rates and significant measurement delays. On the other hand, online-quality instruments tend to be costly and require complex maintenance. The real-time process monitoring, control, and optimization requirements are not sufficiently satisfied by either offline tests or online sensors. Soft sensors are increasingly being used to quickly estimate process parameters in real-time due to their rapid reactivity, maintenance-cost effectiveness, and accurate prediction results. Soft sensors can forecast difficult-to-measure quality metrics by developing mathematical models based on secondary process parameters that are straightforward to monitor, such as temperatures, pressures, and flow rates (Shao and Tian, 2015).

Artificial neural networks (ANNs) or advanced multiple regression techniques provide an alternative approach for acquiring values of process variables that are not directly measurable. These estimators are typically called "software sensors", emphasizing their significance in providing process variable values as if they were generated by physical sensors intended for monitoring and control applications. Researchers have recently employed data-driven-based soft sensor approaches such as nonlinear regression and Artificial Neural Networks (ANNs) to estimate the biomass.

However, they have some limitations such as limited applicability for highly nonlinear systems, inefficient utilization of unlabeled data, inability to predict multiple outputs concurrently, and demand for extensive computational resources to handle substantial data (Havlik et al., 2022). Deep neural networks (DNN) proposed byHinton et al. (2007) can successfully overcome the challenges associated with network training by employing a combination of layer-wise unsupervised pretraining and supervised fine-tuning. Deep architectures exhibit superior generalization when dealing with highly different nonlinear functions since they comprise several layers of parameterized nonlinear features (Bengio et al., 2005).

Rao et al. (2023) proposed extreme gradient boosting (XGBoost) soft sensors, utilizing surrogate indicators to emulate algal cell density (ACD). This study chose seven crucial elements from indicators associated with water quality, meteorology, and temporal factors and then developed the model using the XGBoost algorithm. Nevertheless, the study attained a predictive performance of 0.76, as XGBoost is not specifically designed for handling time series data. Wang et al. (2023) proposed a soft sensor utilizing a four-input (ANN) model, with input parameters such as fermentation time, dissolved oxygen concentration, initial glucose concentration, and added quantities of sodium hydroxide. The ANN-based soft sensor enabled accurate online monitoring of dry cell weight, glucose concentration, and lipid production with high accuracy. Nevertheless, shallow ANNs have a limited capacity to represent complicated functions, and their generalization capability is constrained to larger systems. Multilayer networks are susceptible to issues related to gradient vanishing and exploding, which can impact their stability and performance. Deep learning-based architecture, such as Long short-term memory, remains unexplored in biotechnological processes.

This paper investigates and compares the efficiency and effectiveness of support vector regression (SVR) and Long Short-Term Memory (LSTM) techniques to predict the growth of microalgae Nannochloropsis sp., in a closed indoor vertical photobioreactor system focusing on a combination of process parameters and optical monitoring features. We propose an LSTM-based soft sensor to estimate the biomass concentration in comparison with SVR based on the evaluation of the model performance using R2 and RSME performance metrics. LSTMs are best suited for time series data because of their ability to capture complex patterns and adapt to changing sequences. The inherent flexibility and adaptability of LSTM contribute to enhanced generalization and the handling of temporal data dynamics. We hypothesize that the LSTM model, leveraging its enhanced capability to capture temporal dependencies, will demonstrate superior performance in predicting microalgae growth compared to the SVR model.

* 1. Methodology
		1. Data Collection and Preprocessing

The dataset employed in this research pertains to a 21 day production campaign of *Nannochloropsis* sp. in a 6 L capacity ProviAPT vertical flat panel array reactor (Roef et al., 2012), within a fully automated lab unit integrating Arduino and Raspberry Pi modules, and controlled by the OpenSCADA system developed by Roman Savochenko (Merchán et al., 2017). Data from the photobioreactor were extracted from August 18, 2023, to September 7, 2023, at 5-second intervals. The campaign followed a semi-batch cultivation strategy involving one daily harvest and concurrent feeding session. The microalgae production operates in a range of 2.00 to 3.00 dry weight grams of algae per liter. The major growth parameters that have undergone experimental changes during the cultivation period include intensity and spectrometrical composition of photosynthetically active radiation, CO2 concentration, and the respective duration of day-night cycles. Our target variable ‘algae dry weight’ is measured offline on four samples per day, with samples being taken at the beginning and at the end of the night cycle, and just prior to and after the harvest/feeding routine. Later, the offline dry-weight data samples in four data points per day are resampled to five-second intervals using linear interpolation. The final cumulative data set comprises 13 input features and algae dry weight measurement as one target variable with datetime as an index (Table 1). After filling in missing values by spline interpolation, the data was separated into training and testing sets before undergoing additional preprocessing steps. The training set comprises 70% of the data, while the remaining 30% is allocated for evaluation. The input sequence for both SVM and LSTM is generated following a sequence of preprocessing steps. The data, initially captured in five-second intervals, is transformed into one-minute intervals by averaging over 12-time windows after applying an exponentially weighted moving average for smoothening (Yu et al., 2020). Input features are further simplified, and the multicollinearity problem is addressed by applying Principal Component Analysis (PCA) and selecting the first five significant components as input features (Sulaiman et al., 2021).

Table 1- The description of the inputs and outputs employed in this study

|  |  |
| --- | --- |
| **Description of the variables used** | **Unit** |
| Unix time point in date time format | Date time |
| Cumulative absorption rate | % |
| Absorption rate of blue light | % |
| Absorption rate of red light | % |
| Absorption rate of green light | % |
| CO2 concentration set point | % |
| Percentage of the intensity of blue light (quality) | % |
| Percentage of the intensity of red light (quality) | % |
| Percentage of the intensity of white light (quality) | % |
| Max CO2 into the system | % |
| Reactor temperature | °C |
| pH value |  |
| Photosynthetic active radiation | µmol \* m-2 \* s-1 |
| Algae dry weight | g |

* + 1. Overview of Support vector regression and Long short-term memory

This paper compared the SVR with the LSTM to predict the biomass concentration in an indoor laboratory-scale reactor. SVR can handle linear and non-linear models by determining the best fitting line (hyperplane) in linear cases, which gathers most data points inside a specified epsilon tube (Cristianini and Scholkopf, 2002). SVR employs various kernels, such as Gaussian and radial basis kernels, to transform the data into a higher-dimensional space for nonlinear problems. This procedure makes it easier to understand the relationships between inputs and outputs. On the other hand, an LSTM, categorized as a type of recurrent neural network (RNN), is particularly good for handling time series (Reddy and Prasad, 2018). The input gate controls the addition of new information to the memory, the forget gate decides whether to keep or discard past data, and the output gate determines the appropriate information for producing the output. As a result, sigmoid functions are employed to regulate the gates for reading, writing, and clearing the memory.

* + 1. Implementation details

The details of the hyperparameters used in this study for training SVR and LSTM models are listed in Table 2.

Table 2- The list of the hyperparameters employed in this study

|  |  |
| --- | --- |
|  **Hyperparameters** | **Values** |
| LSTM hidden units | 50 |
| Number of layers | 3 |
| SVR radial basis function | 1 |
| Batch size | 100 |
| Number of epochs | 100 |
| Window size | 3 |
| Learning rate | 0.0001 |
| Optimizer | Adam |
| Library | Python (PyTorch) |

* 1. Results and Discussions

The results in Table 3 reveal that the LSTM model outperforms the SVR model in biomass prediction for the testing set, demonstrating a better fit evidenced by a higher R2 score and lower RMSE as shown in Figure 1. However, the SVR model shows signs of overfitting, as indicated by its superior performance on the training dataset compared to the testing dataset as shown in Figure 1. The disparity in performance between the LSTM and SVR models can be attributed to their inherent algorithmic characteristics. LSTM, a type of RNN, excels at capturing temporal dependencies through its memory cells, making it particularly adept at handling time series data. It can express complex patterns and adapt to different sequences, thus offering better generalization, which is reflected in its superior R2 value. In contrast, nonlinear SVR is a kernel-based technique that transforms an input into a higher-dimensional space using fixed-width kernels. This approach often struggles with capturing complex temporal relationships in time series data, leading to overfitting when the model tries to fit noise instead of meaningful patterns.

**The findings of this study indicate that the LSTM model demonstrated remarkable accuracy in predicting biomass dry weights ranging from 0 to 3 grams per liter, aligning with the Proviron production strategy's operational range. As illustrated in Figure 1, the predictive accuracy diminishes for higher dry weight measurements, particularly within the 2.5 g/L to 3 g/L biomass dry weight range. Further validation is required, and addressing this issue may involve training the model on a broader range of biomass concentrations.** The LSTM model, trained on Nannochloropsis cultivation data, surpasses the performance of the SVR model, attaining a higher R2 score of 0.91 compared to R2 score of SVR 0.781. Despite its overall efficacy, the LSTM model sometimes encounters difficulties with certain instances or patterns, resulting in occasional large offsets as shown in Figure 1. This could be due to its sensitivity to specific data properties, outliers, or challenges in generalizing certain patterns, especially in learning from feed and harvest instances. While LSTM generally provides more accurate predictions, these occasional large offsets can be significant in practical applications. SVRs, effective at handling linear separations, captured the influence of feed and harvest instances represented in Boolean format. In contrast, LSTM, though superior to sequential data, sometimes fails to capture underlying patterns associated with these instances. **Further validation is needed to overcome the large offset encountered in the LSTM.**





Figure 1 – Prediction performance of SVR and LSTM model on the test data

Table 3- R2 and RMSE of LSTM and the SVR for the predictions of the biomass concentration

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Criterion** | **LSTM** | **SVR** |
| Train | R2 | 0.991 | 0.929 |
| RMSE | 0.042 | 0.088 |
| Test | R2 | 0.91 | 0.781 |
| RMSE | 0.061 | 0.723 |

* 1. Conclusion

The results of this study confirm that the use of machine learning models such as SVR and LSTM can predict microalgae growth by incorporating biomass concentration or optical density and time series input features. Specifically, the LSTM model demonstrates

a remarkable capability in capturing the inherent complex behavior of the process better than the SVR model. In summary, this study introduces a soft sensor utilizing LSTM for predicting microalgae growth, overcoming issues in the online measurement of essential biological process parameters. This enhancement is significant for real-time monitoring and optimizing microalgal biomass concentration.

References

W. Shao, X. Tian, 2015, Adaptive soft sensor for quality prediction of chemical processes based on selective ensemble of local partial least squares models, Chem. Eng. Res. Des. 95,113-132.

W. Rao, X. Qian, Y. Fan, T. Liu, 2023, A soft sensor for simulating algal cell density based on dynamic response to environmental changes in a eutrophic shallow lake, Sci. Total Environ. 868, 161543.

K. Wang, W. Zhao, L. Lin, T. Wang, P. Wei, R. Ledesma-Amaro, A. Zhang, X.Ji, 2023, A robust soft sensor based on artificial neural network for monitoring microbial lipid fermentation processes using Yarrowia lipolytica, Biotechnol. Bioeng, 120, 1015–1025.

G.E. Hinton, 2007, Learning multiple layers of representation, Trends Cogn. Sci. 11, 428–434.

Y. Bengio, O. Delalleau, N. Roux, 2005, The Curse of Highly Variable Functions for Local Kernel Machines, in: Advances in Neural Information Processing Systems, MIT Press.

I. Havlik, S. Beutel, T. Scheper, K.F. Reardon, 2022, On-Line Monitoring of Biological Parameters in Microalgal Bioprocesses Using Optical Methods, Energies 15, 875.

M.S. Sulaiman, M.M. Abood, S.K. Sinnakaudan, M.R. Shukor, G.Q. You, X.Z. Chung, 2021, Assessing and solving multicollinearity in sediment transport prediction models using principal component analysis, ISH J. Hydraul. Eng. 27, 343–353.

J. Yu, S.B. Kim, J. Bai, S.W. Han, 2020, Comparative Study on Exponentially Weighted Moving Average Approaches for the Self-Starting Forecasting, Appl. Sci. 10, 7351.

N. Cristianini, B. Scholkopf. 2002, Support Vector Machines and Kernel Methods: The New Generation of Learning Machines, AI Mag. 23, 31–31.

D.S. Reddy, P.R.C. Prasad, 2018, Prediction of vegetation dynamics using NDVI time series data and LSTM, Model. Earth Syst. Environ. 4, 409–419.

L. Roef, M, Jacqmain, M. Michiels, 2012, 13 Case study: Microalgae production in the selfsupported ProviAPT vertical flat-panel photobioreactor system, in: Posten, C., Walter, C. (Eds.), Microalgal Biotechnology: Potential and Production, DE GRUYTER, pp. 243–246.

D.F. Merchan, J.A. Peralta, A. Vazquez-Rodas, L.I. Minchala, D. Astudillo-Salinas, 2017, November, Open source SCADA system for advanced monitoring of industrial processesIEEE International Conference on Information Systems and Computer Science (INCISCOS), pp. 160-165.