Scientific Papers. Series E. Land Reclamation, Earth Observation & Surveying, Environmental Engineering. Vol. XIII, 2024
Print ISSN 2285-6064, CD-ROM ISSN 2285-6072, Online ISSN 2393-5138, ISSN-L 2285-6064

ENABLING SOFT SENSORS FOR WATER QUALITY MONITORING IN MULTI-TROPHIC AQUACULTURE SYSTEMS

Stefan-Mihai PETREA[1, 2, 4], Ira-Adeline SIMIONOV[1, 2, 5], Alina ANTACHE[1, 2, 6], Aurelia NICA[1], Madalina CALMUC[2, 3], Valentina CALMUC[2, 3], Dragos CRISTEA[4], Puiu Lucian GEORGESCU[2, 3]

[1]"Dunarea de Jos" University of Galati, Food Science and Engineering Faculty, 111 Domneasca Street, Galati, Romania
[2]"Dunarea de Jos" University of Galati, REXDAN Research Infrastructure, 98 George Cosbuc Street, Galati, Romania
[3]"Dunărea de Jos" University of Galati, Faculty of Sciences and Environment, 111 Domneasca Street, Galati, Romania
[4]"Dunărea de Jos" University of Galati, Faculty of Economics and Business Administration, 59-61 Nicolae Balcescu Street, Galati, Romania
[5]"Dunărea de Jos" University of Galati, Faculty of Automatic Control, Computers, Electrical Engineering and Electronics, 2 Stiintei Street, Galati, Romania
[6]"Alexandru Ioan Cuza" University of Iasi, 11 Carol I Blvd, Iasi, Romania

Corresponding author email: stefan.petrea@ugal.ro

Abstract

The paper aims to use machine-learning-based algorithms in order to enable and empower the integration of soft sensors for improving the economic sustainability of integrated multi-trophic recirculating aquaculture systems (IMRAS) through efficient and accurate water quality monitoring of nitrate (NO_3), the main key parameter for maintaining the sustainability of the IMRAS in various production scenarios. A 30-day trial was conducted in a sturgeon–tarragon IMRAS to develop a NO_3 soft sensor, based on a series of predictors such as pH, temperature, NH_4, NO_2, NO_3, conductivity (EC), P_2O_5, Ca and Mg, as well as to identify the prediction model peculiarities in various exploitation scenarios generated by the crops culture density. The results reveal the effectiveness of different learning algorithms as MLR and XGBoost (>80% accuracy) in developing solutions for supporting the water quality monitoring process in IMRASs, concluding that the intensity of production technologies must be considered as a determinant factor in upscaling the solutions to industrial level.

***Key words**: soft sensor, machine-learning, nitrate, water quality, multi-trophic aquaculture.*

INTRODUCTION

The European Committee (EC) policy strategies, through its plan of transition from a Blue to a Green economy, emphasize the strategic orientation towards sustainable and competitive aquaculture within the European Union (EU) borders, which is targeted to be implemented within the period 2021-2030.
In order to support this desideratum, research oriented towards the identification of scalable and replicable innovative technical and technological solutions, based on artificial intelligence, which are able to support the maximization of sustainability degree associated with the emergent integrated production technologies, a multi-disciplinary analytical framework must be considered and research niches that can improve various peculiar processes with the technology must be tackled.
Since aquaculture is a major part of Blue Economic - Development, the shift toward a Green Economy imposes the extent of sustainable aquaculture systems and practices, limiting therefore their negative impact on the environment, mostly associated with the degree of production intensity. Thus, previous studies (Paepae et al., 2021; Petrea et al., 2023a) present integrated aquaponics as a feasible solution for limiting aquaculture wastes by using circular economy principles that consist of valorizing the wastes for obtaining a second crop production that can also be commercialized and, therefore, can contribute to the increase of economic

competitiveness, a fact confirmed by Bosma et al. (2017), Petrea et al. (2019), Asciuto at al. (2019) and Costache et al. (2021).

However, integrating aquaponics into already existing aquaculture conventional or recirculating aquaculture systems-based farms is a complex process, both from technical and technological perspectives, as resulting from various studies (Yildiz et al., 2017; Goddek et al., 2019). The most important process within an aquaponic system is related to nutrient management – thus, continuous water quality monitoring is imposed to limit potential dysfunctionalities which can have fatal impacts both for fish and plant production.

According to a previous study (Petrea et al., 2023b), nitrate is considered one of the most important water quality parameters in aquaculture systems and has an accumulation trend – this situation can be efficiently exploited by aquaponics technologies since nitrate is an essential nutrient during the plants culture process.

Petrea et al. (2023a) confirmed that soft sensors can be successfully used in aquaponics multi-trophic systems in order to accurately predict the concentration of essential water quality parameters.

However, since each integrated aquaponic technology has its peculiarities related to various variables such as plant and fish species used, technology production intensity (e.g. feed input, fish stocking density, plant production density etc.) or aquaponics techniques used (e.g. floating rafts, nutrient film technique, substrate techniques etc.), the develop of artificial intelligence (AI) - based soft sensors cannot have yet, a universal applicability.

Thus, if the generative character of AI will be exploited in future, considering the soft sensors' analytical frameworks, this limitation will be overcome.

The present research targets to apply machine-learning-based algorithms in order to enable and empower the integration of soft sensors for improving the economic sustainability of integrated multi-trophic recirculating aquaculture systems (IMRAS) through efficient and accurate water quality monitoring of nitrate, considering a series of predictors, specific for IMRAS and various technological scenarios.

MATERIALS AND METHODS

Experimental design, data collection and dataset description

In order to create the dataset used for developing the soft sensor for determining water NO_3 concentration in an aquaponic system, a 30-day trial was conducted in a sturgeon (*A. baeri*) – tarragon (*Artemisia dracunculus* L.) IMRAS where light expanded clay aggregate (LECA) aquaponic substrate was used as plant growing substrate and 3 different tarragon culture densities were applied, together with the control variant, as follows:

- *D1* – culture density of tarragon - 80 plants/m^2;
- *D2* – culture density of tarragon - 60 plants/m^2;
- *D3* – culture density of tarragon - 40 plants/m^2;
- *C* – Control variant – no plants were used, only LECA plants' culture substrate.

During the experimental trial, the following water quality parameters concentrations were monitored: NH_4 (ammonium), NO_2 (nitrites), NO_3 (nitrates), EC (electroconductivity), P_2O_5 (phosphorus pentoxide), Ca (calcium) and Mg (magnesium). Thus, all parameters were determined using Libelium® Smart Water Sensor Platform Adds Ion Monitoring (Zaragoza, Spain) - the equipment is fully described by Petrea et al., 2023a. However, P_2O_5 was the only parameter determined by using laboratory analytical procedures – sensors were calibrated before the trial and validation of the data was performed by crosschecking the sensor results with the results obtained by applying the classical analytical determination procedures. The sampling points were both at the inlet (*In*) and outlet (*Out*) of the aquaponics modules, for each experimental variant, to create a decision-support analytical framework, oriented towards NO_3, that can be developed, further on, in order to serve as a solution for determining the real-time NO_3 removal rate.

The resulting dataset was divided into 2 groups, as follows: the 1st group contains 70% of the data allocated for training, while the 2nd group has the rest 30% of data, that are to be used for the validation process. Database pre-processing is described by Petrea et al. (2023a), while the predictor's standardization was performed as presented by Petrea et al. (2020).

Machine learning data-processing methods and workflow

In order to perform the analytical framework of the present study, the workflow diagram described in Figure 1 was applied.

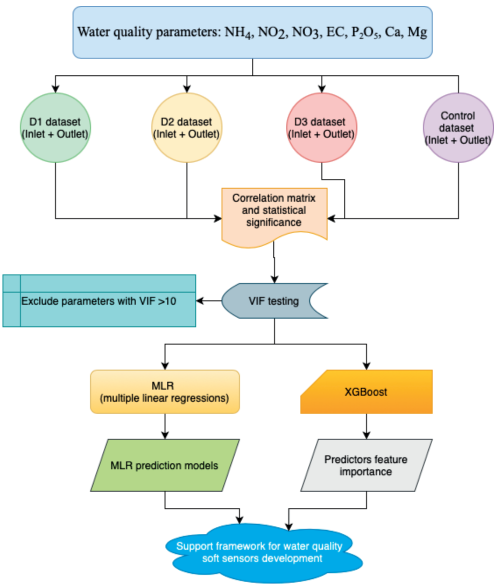

Figure 1. The workflow for enabling and empowering the integration of soft sensors in IMRAS, considering various technological scenarios

Thus, a number of 2 machine learning-based supervised algorithms were used, namely multiple linear regression (MLR) and XGBoost (XGB), in order to generate high-metrics predictions of NO_3 concentration in water, considering 6 main predictors, associated with each of the 4 technological scenarios (*D1, D2, D3, C*), both at the inlet (*In*) and outlet (*Out*) of the aquaponic units.

The Python NumPy library was used for obtaining the correlation matrix and Seaborn library to visualize it, as presented by Petrea et al., 2023b.

The MLR model equation is presented below (eq. 1), according to Petrea et al. (2023a):

$$y_{predict} = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b + \epsilon \quad (1)$$

where:

- $y_{predict}$ is the dependent variable;
- x_1, $x_2 \ldots x_n$ are the n independent variables;
- b is the intercept indicating the Y value when all the predictors are zeros;
- a_1, a_2 ... a_n are the coefficients of predictors, reflecting the contribution of each independent variable in predicting the dependent variable; ε is the residual term indicating the difference between the actual and the fitted response value.

The XGBoost core-concept is presented in eq. 2.

$$Obj = \sum_{i=1}^{n} l\left(y_i - \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad (2)$$

where:

- y_i is the measured value; y
- i is the predicted value of each tree; l is the loss function, which is used to measure the total prediction error;
- $\sum\Omega(fk)$ is the regularization term.

RESULTS AND DISCUSSIONS

Correlation matrix

The matrixes reveal that, in the case of D1, the highest significant ($p<0.05$) positive correlations are recorded between the pH and EC, P_2O_5 and NO_3, respectively, while the significant ($p<0.05$) high negative correlations are observed between EC – P_2O_5, P_2O_5 – NO_3 and EC – NO_3 (Figure 2).

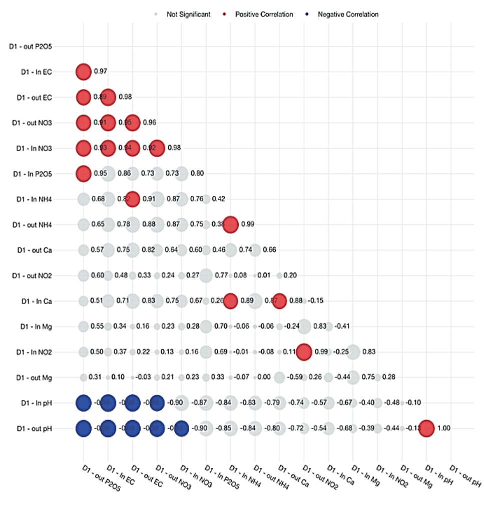

Figure 2. Correlation matrix for water quality parameter dataset associated with *D1* technological scenario

In the case of D2, the highest significant ($p<0.05$) positive correlations encountered at D1 are confirmed, the pH dynamics conditioning the EC, P_2O_5 and NO_3, fact valid both for *In* and *Out* sampling points (Figure 3). In terms of significant ($p<0.05$) high negative correlations, outside EC – P_2O_5 – NO_3 conditioning triangle, Ca – NH_4 and Mg – NO_3 relations are also pointed out as negative and significant ($p<0.05$) (Figure 3).

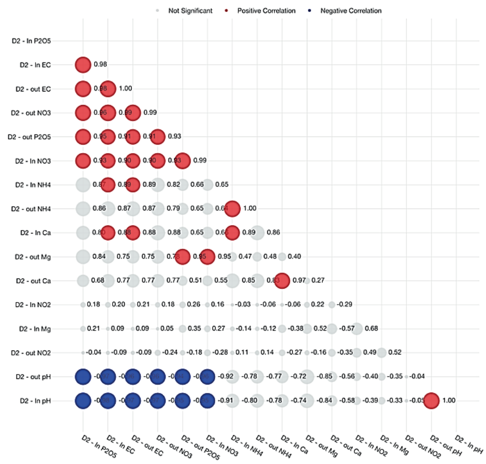

Figure 3. Correlation matrix for water quality parameter dataset associated with *D2* technological scenario

The matrixes in the case of D3 reveal the highest significant ($p<0.05$) positive correlations recorded between the pH and EC, P_2O_5 and NO_3, while the significant ($p<0.05$) high negative correlations are observed between EC – P_2O_5 – NO_3 – NH_4 conditioning nexus, Ca- EC and Mg, NO_2 and pH *In* – *Out* (Figure 4).

In the case of the C experimental variant correlation matrix (Figure 5), the highest significant ($p<0.05$) positive correlations were recorded, similar to D1, D2 and D3, between the pH and EC, P_2O_5 and NO_3, while the significant ($p<0.05$) high negative correlations are observed between EC, NH_4, P_2O_5, Mg, Ca In – Out, EC – NO_3, EC-NH_4, Ca – NO_3 and P_2O_5 – NO_3.

It can be observed that D1, D2 and D3 experimental variants present less correlation between the concentration of the parameter in *In* vs. *Out* sampling points, respectively, compared to C, a fact that can be due to tarragon dynamics to absorb nutrients that cannot be structured in a pattern based on correlation matrixes. Thus, to perform an in-depth analysis of the dataset conditionalities as a base tool for future development of soft sensors, AI-based machine learning algorithms must be applied.

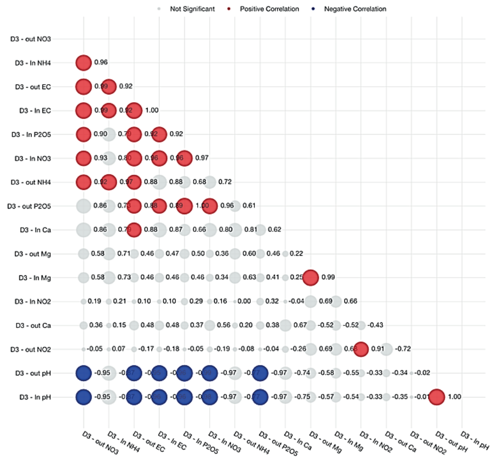

Figure 4. Correlation matrix for water quality parameter dataset associated with *D3* technological scenario

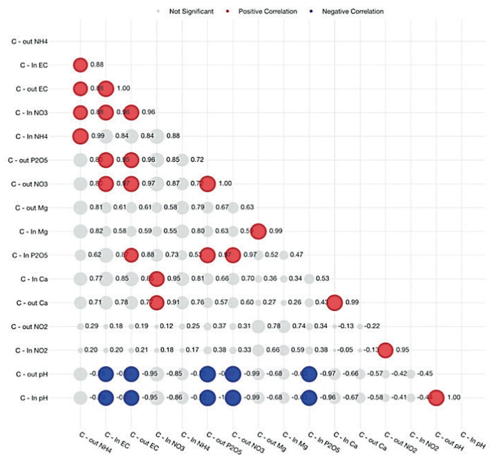

Figure 5. Correlation matrix for water quality parameter dataset associated with *D4* technological scenario

MLR prediction models

In the case of D1, the MLR-based prediction models reveal that NH_4 and Mg can be considered the main predictors for determining the NO_3 concentration in IMRAS technological water, both for *In* (eq. 3) and *Out* (eq. 4) sampling points, since they are associated with the highest coefficient values within the model.

All parameters were used as predictors since their VIF values were lower than 10.
The models' metrics reveal good accuracy, with an Rsq value of 83.24 for the $NO_{3\ D1in}$ prediction model (eq. 3) and 80.45 for the $NO_{3\ D1out}$ prediction model (eq. 4). However, a positive predictive relation can be observed between NO_3 and the predictors EC and NH_4, while negative predictive weight is associated to Ca and Mg, a fact valid for both *In* and *Out* sampling points.

- $NO_{3\ D1in} = 1059.48 - 30.69\ Ca_{D1in} + 2.75\ EC_{D1in} - 46.96\ Mg_{D1in} + 1325.72\ NH_{4\ D1in}$ (3)
- $NO_{3\ D1out} = 951.23 - 25.92\ Ca_{D1out} + 2.25\ EC_{D1out} - 38.93\ Mg_{D1out} + 1301.42\ NH_{4\ D1out}$ (4)

In the case of D2, the MLR-based prediction models reveal that most of the NO_3 dynamics can be explained by NH_4, a fact valid for both *In* and *Out* sampling points (eq.5 and eq.6). However, the Ca and Mg predicting weight increases in the case of $NO_{3\ D2in}$ (eq.5), compared to $NO_{3\ D2out}$ (eq.6). All parameters were used as predictors since their VIF values were lower than 10. The models' metrics reveal good accuracy, with an Rsq value of 86.12 for the $NO_{3\ D2in}$ prediction model (eq.5) and 81.87 for the $NO_{3\ D2out}$ prediction model (eq.6).

- $NO_{3\ D2in} = -632.16 - 12.55\ Ca_{D2in} + 2.12\ EC_{D2in} - 16.5\ Mg_{D2in} - 199.77\ NH_{4\ D2in}$ (5)
- $NO_{3\ D2out} = -422.30 - 1.38\ Ca_{D2out} + 0.83\ EC_{D2out} - 3.28\ Mg_{D2out} - 157.13\ NH_{4\ D2out}$ (6)

In the case of D3, the MLR-based prediction models reveal that NO_3 prediction is mostly conditioned by NH_4 concentration in the technological water, in both *In* and *Out* sampling points (eq. 7 and eq. 8). However, the $NO_{3\ D3in}$ model emphasizes the low weight of Ca in the prediction equation (eq. 7), compared to $NO_{3\ D3out}$. Also, it can be concluded that both EC and Mg share a similar trend of predictive weight (eq. 7 and eq. 8). The models' metrics reveal good accuracy, with an Rsq value of 83.34 for the $NO_{3\ D3in}$ prediction model (eq.7) and 80.52 for the $NO_{3\ D3out}$ prediction model (eq.8). The NH_4 has a direct trend in relation to the predicted variable, in the case of *Out* sampling point, while in the *In* sampling point, the relation is indirect (negative). This can be explained by the autoregulation capacity of the aquaponic unit, in terms of the water quality matrix.

- $NO_{3\ D3in} = -585.46 + 0.06\ Ca_{D3in} + 0.77\ EC_{D3in} + 3.05\ Mg_{D3in} - 421.42\ NH_{4\ D3in}$ (7)
- $NO_{3\ D3Out} = -239.95 + 0.55\ Ca_{D3Out} + 0.41\ EC_{D3Out} + 1.86\ Mg_{D3Out} + 33.54\ NH_{4\ D3Out}$ (8)

The C experimental variant presents good accuracy metrics, with an Rsq value of 89.58 for the $NO_{3\ Cin}$ prediction model (eq. 9) and 84.32 for the $NO_{3\ Cout}$ prediction model (eq. 10). In both cases, the NH_4 has the highest weight in predicting the NO_3, followed by Ca and Mg. However, it can be pointed out that Ca has an indirect trend in relation to the predicted variable, in the case of *Out* sampling point, while in the *In* sampling point, the relation is direct (positive).

- $NO_{3\ Cin} = -318.82 + 4.40\ Ca_{Cin} + 0.18\ EC_{Cin} + 3.49\ Mg_{Cin} - 155.86\ NH_{4\ Cin}$ (9)
- $NO_{3\ C\ out} = -574.48 - 1.83\ Ca_{C\ out} + 0.89\ EC_{C\ out} + 2.41\ Mg_{C\ out} - 241.66\ NH_{4\ C\ out}$ (10)

In the end, it can be partially concluded that the MLR models emphasize that NH_4 has a considerable weight, being recommended to be used as the main predictor for the development of the NO_3 soft sensors, applicable in IMRAS, in various technological scenarios.

XGBoost prediction models

The XGBoost prediction models, applied by using the D1 dataset, reveal that for *In* sampling point (Figure 6), the In-EC and In-Ca are the only predictors that present future importance associated with the prediction of the In-NO_3 concentration in the technological water matrix. However, in the case of the *Out* sampling point (Figure 7), the Ca has the highest future importance, followed by EC.
The metrics of both D1 prediction models (Figures 6, 7) indicate good accuracy, as follows: Rsq of 91.37 for the *In* sampling point and 89.34 for the *Out* sampling point, respectively.
The XGBoost prediction models, applied by using the D2 dataset, reveal that for *In* sampling point (Figure 8), the In-EC, In-P_2O_5 and In-Ca are the only predictors that present future importance in predicting the In-NO_3 concentration at the level of technological water. However, in the case of the *Out* sampling point (Figure 9), the Ca has the highest future importance, followed by EC, while the P_2O_5 has no associated value for feature importance.

Thus, it can be concluded that the variation of tarragon culture density from 80 plants/m^2 to 60 plants/m^2 has an impact on predictors feature importance when predicting the NO_3 concentration from aquaponics modules inlet sampling point, recommending, therefore, to consider P_2O_5 among EC and Ca, as main predictors.

The metrics of both D2 prediction models (Figures 8, 9) indicate good accuracy, better than D1 models, as follows: Rsq of 93.12 for the *In* sampling point and 90.44 for the *Out* sampling point, respectively.

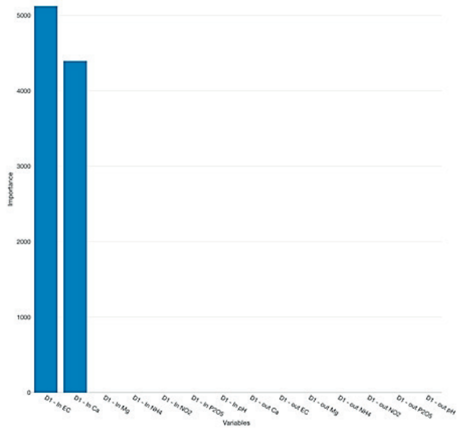

Figure 6. The predictors feature importance in predicting NO_3 concentration at D1-In

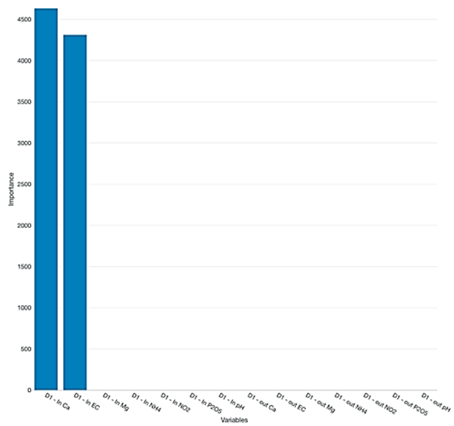

Figure 7. The predictors feature importance in predicting NO_3 concentration at D1-Out

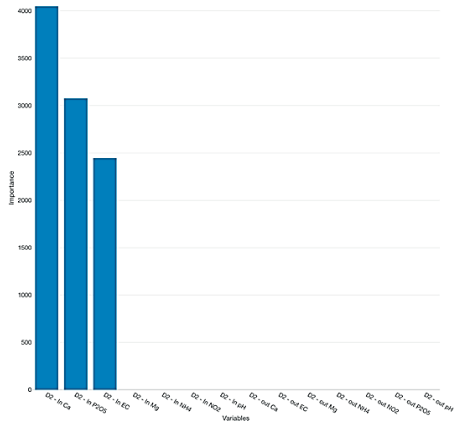

Figure 8. The predictors feature importance in predicting NO_3 concentration at D2-In

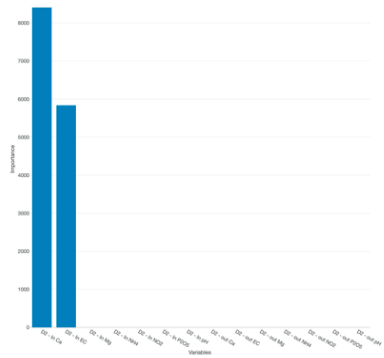

Figure 9. The predictors feature importance in predicting NO_3 concentration at D2-Out

The NO_3 - XGBoost prediction models applied by using the D3 dataset reveal that, for both sampling points, the EC is considered as the main predictor, followed by Ca (Figures 10, 11). Thus, the findings confirm, generally, the predictors ranking resulted from MLR models, presented above. However, the metrics of both D3 prediction models (Figures 10, 11) indicate good accuracy, better than D1 and D2 models, as follows: Rsq of 95.55 for the *In* sampling point and 93.12 for the *Out* sampling point, respectively.

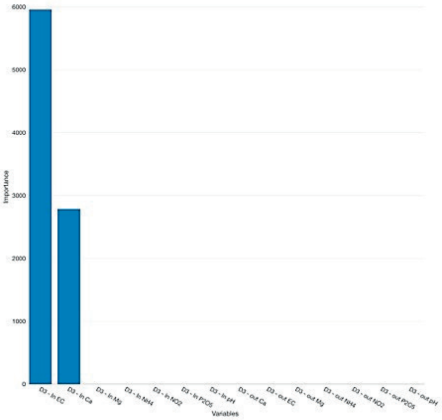

Figure 10. The predictors feature importance in predicting NO_3 concentration at D3-In

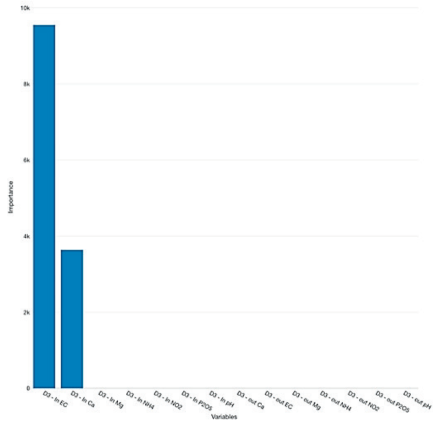

Figure 11. The predictors feature importance in predicting NO_3 concentration at D3-Out

The C experimental variant confirms the Ca feature importance, as a predictor, in predicting the NO_3 concentration of technological water, in the *In* sampling point (Figure 12). However, at the *Out* sampling point, the situation is different, with the EC recording the highest feature importance among the predictors (Figure 13). Also, compared to *In* sampling point, the model for predicting the NO_3 concentration in water at the aquaponics units outlet attributes a feature importance value, also, to P_2O_5, outside EC and Ca.

The metrics of both C prediction models (Figure 12, 13) indicate good accuracy, better than D1, D2 and D3 models, as follows: Rsq of 97.10 for the *In* sampling point and 95.14 for the *Out* sampling point, respectively. This can be explained by the lack of plant biomass at the level of control variant aquaponics modules – thus, this makes the model more stable and emphasizes the complexity degree in identifying high-accuracy models for various aquaponics culture density scenarios.

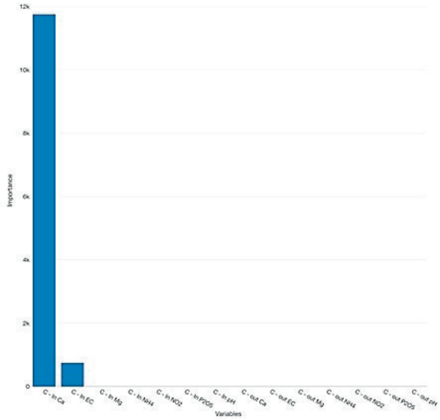

Figure 12. The predictors feature importance in predicting NO_3 concentration at C-In

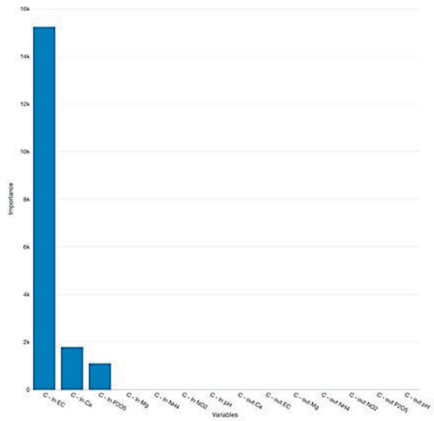

Figure 13. The predictors feature importance in predicting NO_3 concentration at C-Out

The metrics of both C prediction models (Figures 12, 13) indicate good accuracy, better than D1, D2 and D3 models, as follows: Rsq of 97.10 for the *In* sampling point and 95.14 for the *Out* sampling point, respectively. This can be explained by the lack of plant biomass at the level of control variant aquaponics modules – thus, this makes the model more stable and emphasizes the complexity degree in identifying

high-accuracy models for various aquaponics culture density scenarios.
Also, it can be concluded that for all experimental variants, the accuracy metrics recorded by applying XGBoost algorithms are superior compared to the metrics resulting because of applying MLR algorithms.
Similar to other studies, this study has its own limitations in terms of dataset dimension, as well as experimental framework. However, the study results (prediction models) are associated with high accuracy metrics, a fact that makes them reliable for upscaling in real industrial conditions.

CONCLUSIONS

The results reveal the effectiveness of different learning algorithms as MLR and XGBoost (>80% accuracy) in developing solutions for supporting the water quality monitoring process in IMRASs, concluding that the intensity of production technologies must be considered as a determinant factor in upscaling the solutions to industrial level.
It is recommended that future research should take into consideration the application of other machine learning algorithms such as GAM, SVM, Random Forest, GBM, CNN, Stacked ensemble, DRF or Naïve Bayesian in order to have a complete background of the technical aspects that may be used for future development of the analytical framework for black-box and grey-box soft sensors.

ACKNOWLEDGEMENTS

The present research was supported by the project: "An Integrated System for the Complex Environmental Research and Monitoring in the Danube River Area", REXDAN, SMIS code 127065, co-financed by the European Regional Development Fund through the Competitiveness Operational Programme 2014–2020, contract no. 309/10.07.2021. The present research was supported by the project "Integrated research and sustainable solutions to protect and restore Lower Danube Basin and coastal Black Sea ecosystems, (ResPonSE)", 760010/30.12.2022, component C9. PRIVATE SECTOR SUPPORT, RESEARCH, DEVELOPMENT AND INNOVATION, Investment "I5. Establishment and operationalization of Competence Centers". (specific project no.1 - Decision Support Solutions based on mathematical modeling, for complex ecological systems – CESMoSS).

REFERENCES

Asciuto, E.S., Cottone, C. & Borsellino, V. (2019) A financial feasibility study of an aquaponic system in a Mediterranean urban context. *Urban For. Urban Green*, 38, 10.1016/j.ufug.2019.02.001.

Bosma, R.H., Lacambra, L., Landstra, Y., Perini, C., Poulie, J., Schwaner, M.J. & Yin, Y. (2017). The financial feasibility of producing fish and vegetables through aquaponics. *Aquac. Eng.*, 78, 10.1016/j.aquaeng.2017.07.002

Costache, M., Cristea, D.S., Petrea, S.-M., Neculita, M., Rahoveanu, M.M.T., Simionov, I.-A., Mogodan, A., Sarpe, D. & Rahoveanu, A.T. (2021). Integrating aquaponics production systems into the Romanian green procurement network. *Land Use Policy*, *108,* 105531.

Goddek, S. & Körner, O.A. (2019). Fully Integrated Simulation Model of Multi-Loop Aquaponics: A Case Study for System Sizing in Different Environments. *Agricultural Systems*, *171,* doi:10.1016/j.agsy.2019.01.010.

Paepae, T., Bokoro, P.N. & Kyamakya, K. (2021). From fully physical to virtual sensing for water quality assessment: A comprehensive review of the relevant state-of-the-art. *Sensors*, *21,* 6971.

Petrea, Ș.-M., Simionov, I.A., Antache, A.; Nica, A., Oprica, L., Miron, A., Zamfir, C.G., Neculiță, M., Dima, M.F. & Cristea, D.S. (2023a). An Analytical Framework on Utilizing Various Integrated Multi-Trophic Scenarios for Basil Production. *Plants, 12*, 540. https://doi.org/10.3390/plants12030540

Petrea, Ș.M., Bandi, A.C., Cristea, D. & Neculiță, M. (2019). Cost-benefit analysis into integrated aquaponics systems. *Custos E Agronegócioon Line*, *15,* 239–269.

Petrea, Ș. -M., Simionov, I. A., Antache, A., Cristea D. S., Nica, A. & Cristea V. (2023b). Analytical framework for the development of water quality virtual sensors in integrated aquaculture systems based on various aquaponics techniques. *IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sinaia, Romania, pp. 1-5. doi: 10.1109/ETFA54631.2023.10275350.

Petrea, Ș.-M., Costache, M., Cristea, D., Strungaru, Ș.-A, Simionov, I.-A., Mogodan A., Oprica, L. & Cristea V. (2020). A Machine Learning Approach in Analyzing Bioaccumulation of Heavy Metals in Turbot Tissues. *Molecules*, *25*, 4696

Yildiz, H.Y., Robaina, L., Pirhonen, J., Mente, E., Domínguez, D. & Parisi, G. (2017). Fish Welfare in Aquaponic Systems: Its Relation to Water Quality with an Emphasis on Feed and Faeces-A Review. *Water (Switzerland)*, *9*.