Automated Machine Learning for Ex-ante Life Cycle Assessment of Barley Production

Radwa ElShawi[1,†], Stefania Tomasiello[1,2,†], Evelin Loit[3], Ayhan Gökçe[4], Virmantas Povilaitis[5], Stefan Shilev[6] and Wieslaw Szulc[7]

[1]*Institute of Computer Science, University of Tartu, Estonia*
[2]*Department of Industrial Engineering, University of Salerno, Fisciano, Italy*
[3]*Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Estonia*
[4]*Niğde Ömer Halisdemir University, Faculty of Agricultural Sciences and Technologies, Türkiye*
[5]*Lithuanian Research Centre for Agriculture and Forestry, Institute of Agriculture, Lithuania*
[6]*Agricultural University, Plovdiv, Bulgaria*
[7]*Warsaw University of Life Sciences, Institute of Agriculture, Poland*

Abstract

In this paper, we discuss the use of Automated Machine Learning for the first time applied to an Ex-ante Life Cycle Assessment. This kind of analysis has been conducted for a particular crop production, i.e. barley. The aim is to assess the impact (in terms of carbon dioxide emissions and yield) of different production strategies. The data used in this study comes from a two-year measurement campaign involving five countries. The results are compared against the state-of-the-art technique, showing the good performance of the approach.

Keywords

automated machine learning, fuzzy logic, Ex-ante Life Cycle Assessment, crop production

1. Introduction

Ex-ante Life Cycle Assessment (LCA), unlike traditional LCA, is conducted before a product, process, or service is fully developed or implemented. Hence, it refers to a predictive or foresight-oriented approach to environmental impact assessment. While traditional LCA assesses the environmental impacts of an existing product or process (post-implementation), ex-ante LCA is performed during the design phase or early planning stages. It aims to forecast potential environmental impacts based on available data and assumptions. It may help decision-makers evaluate different alternatives and make informed choices about which design or process will likely have the least environmental impact. Since ex-ante LCA is conducted early in the product lifecycle, it often relies on estimated data or assumptions that might not be as accurate as data obtained after production or use. This introduces some uncertainty in the analysis, but it still provides valuable insights for comparison. It allows for the early identification of potentially unsustainable practices, guiding toward more sustainable solutions. It brings some benefits, since assessing environmental impacts early in the process can help avoid costly mistakes and optimize the environmental performance of a product or service. The limitations mainly deal with uncertainty, since early-stage assessments are based on models and assumptions that may not fully reflect the actual environmental impact once the product or process is realized [1]. The use of ex-ante LCA in the agri-food area is still in its infancy. In [2], a kind of ex-ante LCA was discussed for the starch extraction process from mango kernel. The goal was to compare the environmental impact of two different processes. The authors used the commercial software SimaPro for their analysis. In [3], ex-ante LCA was used to assess the sustainability of the production of cultivated meat, i.e. non-conventional meat, obtained by processing some crops. The goal was to outline the environmental performance of commercial-scale production and to compare this to conventional meat production in 2030. The authors used the conventional ReCiPe Midpoint impact assessment method for their analysis. In [4], the ex-ante LCA was performed to evaluate the potential changes in fertilizer application rates for a certain wheat production. Machine learning (ML) was not utilized in any of these studies. To the best of our knowledge, the only one in the agri-food area is [5], where the authors proposed the Co-Active Neuro-Fuzzy Inference System with fractional regularization (CANFIS-T) for the ex-ante LCA of wheat production, predicting the yield and the Global Warming Potential (GWP) indicator.

In this work, we explore the integration of an Automated Machine Learning (AutoML) technique to enhance ex-ante LCA, specifically for predicting crop yields and CO2 emissions. AutoML provides an efficient framework for automating the process of model selection and hyperparameter optimization, enabling rapid development of predictive models with minimal manual intervention [6, 7]. Previous studies have demonstrated the potential of AutoML in agricultural applications, such as optimizing models for crop yield prediction and resource management [8, 9]. Similarly, in environmental contexts, AutoML has been used to predict greenhouse gas emissions and assess the sustainability of processes with limited data [10]. These studies highlight the adaptability and scalability of AutoML for addressing complex, data-driven challenges. By leveraging AutoML, we aim to improve the accuracy of early-stage environmental impact predictions while addressing the uncertainties associated with data and assumptions in ex-ante LCA. This approach represents a novel application of AutoML in the agri-food domain, contributing to the advancement of predictive methodologies for sustainable agriculture.

Workshops of the EDBT/ICDT 2025 Joint Conference (March 25-28, 2025), Barcelona, Spain
** Corresponding author.
† These authors contributed equally.
radwa.elshawi@ut.ee (R. ElShawi); stefania.tomasiello@ut.ee (S. Tomasiello); evelin.loit-harro@ut.ee (E. Loit); ayhan.gokce@ohu.edu.tr (A. Gökçe); Virmantas.Povilaitis@lammc.lt (V. Povilaitis); stefan.shilev@au-plovdiv.bg (S. Shilev); wieslaw_szulc@sggw.edu.pl (W. Szulc)
1234-5678-9012 (R. ElShawi); 0000-0001-8208-8285 (S. Tomasiello)

The remainder of this paper is organized as follows. Section 3 introduces the methods and the data collection. Section 4 describes the details of our experimental setup and the evaluation of the proposed methodology, before concluding the paper in Section 5.

2. Ex-ante LCA: an Overview

By looking at the definitions in the literature, the ex-ante LCA aims to [11]:

- "scale up an emerging technology using likely scenarios (e.g., using expert help, extreme views, learning curves for similar technologies) of future performance at full operational scale";
- "compare the emerged technology at scale with the evolved incumbent technology".

Similarly, it can be regarded as an "environmental [LCA] of a new technology before it is commercially implemented in order to guide R& D decisions to make this new technology environmentally competitive as compared to the incumbent technology mix" [12].

The challenging part of conducting an ex-ante LCA is the lack of representative data for the system assessed which could bring uncertainty to the study [12]. In [13], uncertainty has been taken into account using different scenarios in the agri-food context. According to other authors [14], these scenarios can be defined on the basis of discussions with relevant stakeholders. In [5], the uncertainty was taken into account by means of the fuzzy sets formalism, which is part of the adopted model, i.e. CANFIS-T.

3. Methods and Material

3.1. Fields Experiments and Data Collection

Data comes from a two-year measurement campaign, over 2022 through 2023, in the context of the SusCrop ERA-NET funded project ConnectFarms [15]. The field operations were conducted in different countries, namely, Lithuania, Estonia, Poland, Bulgaria, and Turkey. Spring barley (Hordeum vulgare), in different varieties (*Laureate* in Estonia, *Zemela* in Bulgaria, *Fantex* in Poland, *Larende* in Turkey, *Luoke* in Lithuania), was grown in pure stands with a sowing rate of 500 seeds/m^2 (Figure1 shows the fields in two participating countries).

There were two fertilizer treatments: non-fertilized treatment (control) and mineral fertilized treatment N23P30K63. All treatments were in four replications. The plot size was 20 m^2. Spring barley yield was 3,8 t/ha on average in fertilized treatment, which was significantly higher than in non-fertilized control (Figure 2). The highest yields were in Turkey (4,2 t/ha), being the south-most testing site. Most likely irrigation helped to achieve the result. Interestingly, fertilization did not affect the yield. In Bulgaria, there were high yields in fertilized treatments (average 3,8 t/ha), which was almost twice as high as controls. Twice the difference in yields was also observed in Estonia.

In order to build our ML model, the considered variables and their range are shown in Table 1. The yield and carbon dioxide emissions represent the output of our model. The whole dataset, in the ranges shown, has 2,000 samples.

The data was normalized for each variable as $\overline{x} = (x - u)/v$, where u is the mean and v is the standard deviation.

Name	Unit	Range
Seeds	kg	[13,200]
Harrowing	h	[0.75,3]
Harvesting	h	[1,1.9]
Ploughing	h	[0.5,2]
Sowing or planting	h	[0.75,3]
Nitrogen	g/m^2	[0,12]
Phosphate	g/m^2	[0,8]
Potassium	g/m^2	[1.5,15]
Compost	kg/m^2	[0.2,1.2]
Biochar	kg/m^2	[0.75,1]
Yield	kg/m^2	[0.16,0.89]
CO2	ppm	[589,639]

Table 1
Input variables and their range

3.2. AutoML Approach

We leverage Auto-sklearn [16], an advanced AutoML system, to automate both model selection and hyperparameter optimization for the construction of CO2 and crop yield prediction models. Built upon Scikit-Learn [17], Auto-sklearn explores a vast search space that includes 15 classifiers, 14 feature preprocessing options, and 4 data preprocessing strategies, resulting in a highly complex configuration involving 110 hyperparameters. To facilitate algorithm selection and tuning, Auto-sklearn utilizes the SMAC (Sequential Model-based Algorithm Configuration) tool [18]. Moreover, Auto-sklearn enhances model stability and performance by combining the top-performing models into an ensemble using a greedy selection strategy. This ensemble approach starts with an empty set and iteratively incorporates models that maximize validation performance. The model's efficacy is assessed using 5-fold cross-validation, with optimization aimed at minimizing the Root Mean Squared Error (RMSE) to ensure robust performance. Auto-sklearn generates the final ensemble model, with a time budget of 10 minutes allocated for model selection and hyperparameter optimization for the CO2 and yield prediction tasks. This constrained time budget was chosen to balance the need for efficient model development with the computational resources available. Despite the brief time limit, Auto-sklearn can efficiently explore the model search space and produce optimized models that provide robust performance within this timeframe.

3.3. Baseline Model

The state-of-the-art baseline model is the Co-Active Neuro-Fuzzy Inference System with fractional regularization (CANFIS-T) proposed in [5], where it was successfully used to predict the GWP indicator, pursuing an ex-ante LCA of wheat production. In [5], the model had 7 inputs (i.e. field operations, transport, nitrogen, phosphate, manure, biocides, medium voltage) and 3 outputs (i.e. wheat grain, wheat straw, GWP). CANFIS-T generalizes the Adaptive Neuro-Fuzzy Inference System with fractional regularization (ANFIS-T). Both models are based on a multi-layered network architecture to describe the Takagi-Sugeno fuzzy inference system, but while ANFIS models Multi-Input-Single-Output (MISO) systems, CANFIS models Multi-Input-Multiple-Output (MIMO) systems. From both systems, it is possible to extract fuzzy IF-THEN rules. In such rules, the antecedents are linguistic variables that can assume a certain value, i.e., a membership degree to a fuzzy set that

(a)

(b)

Figure 1: Fields in (a) Poland; (b) Turkey.

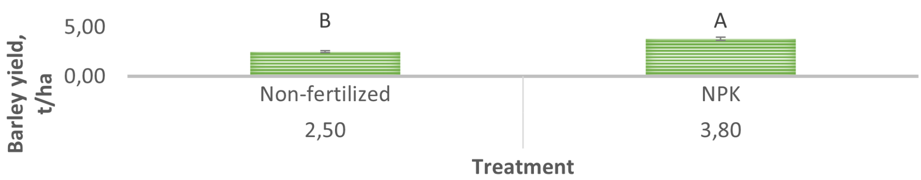

Figure 2: Barley yields in two-year and five-location average ± standard deviation, kg m^2; significant differences are marked with capital letters.

models the variable itself. A fuzzy set is uniquely identified through a so-called membership function, which could be, for instance, a triangular, a trapezoidal, or a Gaussian function. A linguistic variable is associated with a number of terms, modelled by fuzzy sets, which express the attributes of the variable, as in our natural language. Such terms form a fuzzy partition.

4. Experimental Evaluation

4.1. Experimental Setup

Hardware Resources. We conducted our experiments on a CPU environment. The CPU environment runs on Windows 11 Pro 64-bit (10.0, Build 22621) with 16 core Intel(R) Core(TM) i9-10885H Processor @ 2.40GHz,32 GB DIMM memory, and 1000 GB SSD data storage. All the approaches have been implemented in Python.

Treatment Type	MSE (train)	RMSE (train)	MSE (test)	RMSE (test)	R^2 Score (test)
Biochar Treatment	-11.9693	3.4586	10.8 753	3.2978	0.9589
NPK Treatment	-0.0073	0.0848	0.0053	0.0728	1.0000
No Fertilizer Treatment	-0.0041	0.0627	0.0029	0.0542	1.0000

Table 2
Performance metrics of AutoML models for predicting CO2 emissions across different treatment types, highlighting mean squared error (MSE), root mean square error (RMSE), and R^2 scores from five-fold cross-validation (average values).

Treatment Type	MSE (train)	RMSE (train)	MSE (test)	RMSE (test)	R^2 Score (test)
Biochar	-0.0504	0.2245	0.0543	0.2331	0.8956
NPK	0.2322	0.4819	0.2185	0.4674	0.9481
No Fertilizer	0.0925	0.3042	0.0938	0.3063	0.8593

Table 3

Performance metrics of yield prediction models for different treatment types, including mean squared error (MSE), root mean square error (RMSE), and R^2 scores derived from five-fold cross-validation (average values).

4.2. AutoML Results of CO2 Emission Prediction

The results of the AutoML models applied for predicting CO2 emissions reveal significant variances in performance across different treatment types, as shown in Table 2. The Biochar Treatment Model achieved a mean R^2 score of 0.9589, indicating that it explains approximately 95.9% of the test set's variance. Despite this strong correlation, the model demonstrated a root mean square error (RMSE) of 3.2978, suggesting that while the model is reliable, there may be underlying complexities in the biochar data or the target variable distribution contributing to the prediction errors. In contrast, both the NPK Treatment Model and the No Fertilizer Treatment Model reached a perfect R^2 score of 1.0000, with RMSE values of 0.0728 and 0.0542, respectively. These results imply that the models accurately predict CO2 emissions in these scenarios without any observable error, highlighting their robustness. The exceptional performance of the NPK and No Fertilizer models suggests that these treatment types exhibit more consistent patterns in CO2 emissions compared to the biochar treatment. Overall, these findings show the effectiveness of the AutoML approach in modelling carbon emissions under varying agricultural practices, with implications for optimizing treatments aimed at reducing greenhouse gas emissions.

4.3. AutoML Results of Yield Prediction

The results of the yield prediction models indicate varying levels of performance across the different treatment types, as shown in Table 3. The Biochar model produced a mean R^2 score of 0.8956, reflecting that approximately 89.6% of the variance in the test set can be explained by the model. This model exhibited a root mean square error (RMSE) of 0.2331, suggesting some level of prediction error, despite the robust performance indicated by its mean squared error (MSE) of 0.0543. In contrast, the NPK model demonstrated a higher R^2 score of 0.9481, meaning it accounts for about 94.8% of the variance in yield, with an RMSE of 0.4674, indicating slightly larger errors in predictions compared to the Biochar model. This model's mean MSE of 0.2185 is also higher, suggesting that while it performs well, it may be less precise than the Biochar treatment. The No Fertilizer model showed the lowest performance among the three, with an R^2 score of 0.8593, indicating it explains approximately 85.9% of the variance in the test set. The RMSE of 0.3063 indicates notable prediction errors. Collectively, these results highlight the effectiveness of the AutoML approach in predicting agricultural yield, with the NPK treatment showing the highest explanatory power, followed by Biochar and No Fertilizer treatments.

4.4. Baseline Characteristics and Outcome

The study included a comparison against the CANFIS-T method, previously proposed to tackle the same problem in [5]. For yield prediction, the best results were obtained by using 3 terms (and an equal number of rules). The mean test RMSE (with five-fold cross-validation) for each case was as follows:

- No fertilization: 0.8144619
- Biochar: 0.721643
- NPK: 2.1265502

For the CO2 prediction, the results were less favourable, with a mean test RMSE of around 10 for all cases. The average training time for the three cases was less than 2 seconds. Unfortunately, the fast training was the only advantage of the approach.

5. Discussion and Conclusion

When compared with the baseline CANFIS-T method, the AutoML framework substantially outperforms it in terms of predictive accuracy. For example, the AutoML models yielded lower RMSE values (0.2331 for Biochar and 0.4674 for NPK) compared to the CANFIS-T method's results, such as 0.721643 for Biochar and 2.1265502 for NPK. These discrepancies highlight the superior generalization ability of AutoML in accurately modelling agricultural data across different treatment types. Additionally, while the CANFIS-T method demonstrated faster training times, its performance in CO2 emission prediction was less favourable, with an RMSE of approximately 10 across all cases. This further underscores the efficacy of AutoML in addressing complex prediction tasks while maintaining a high level of model stability and accuracy, particularly in cases with higher data variability.

Overall, the findings suggest that AutoML offers significant advantages in both predictive performance and model generalization over traditional methods, providing valuable insights for optimizing agricultural practices aimed at mitigating greenhouse gas emissions.

In future work, we will investigate the use of AutoML techniques to handle uncertainty and noisy data in agriculture.

Acknowledgments

This work has been partially funded by the Estonian Research Council, grant PRG1604, through the funding of SusAn, FACCE ERA-GAS, ICT-AGRI-FOOD and SusCrop ERA-NET, and through the project *Increasing the knowledge intensity of Ida-Viru entrepreneurship* co-funded by the European Union.

References

[1] Tsalidis GA, Korevaar G (2022) Environmental assessments of scales: the effect of ex-ante and ex-post data

on life cycle assessment of wood torrefaction. Res Conser Recyc 176:105906

[2] Pereira da Silva, A.K., et al. Integrating life cycle assessment in early process development stage: The case of extracting starch from mango kernel, Journal of Cleaner Production, 32125 (2021) 128981

[3] Sinke, P., et al. Ex-ante life cycle assessment of commercial-scale cultivated meat production in 2030. Int J Life Cycle Assess 28 (2023) 234–254

[4] Leon A et al (2022) An ex-ante life cycle assessment of wheat with high biological nitrification inhibition capacity. Environ Sci Pollut Res 29:7153–7169

[5] Tomasiello, S., Uzair, M., Liu, Y. et al. Data-driven approaches for sustainable agri-food: coping with sustainability and interpretability. J Ambient Intell Human Comput 14 (2023) 16867–16878

[6] H. Eldeeb, M. Maher, R. Elshawi, and S. Sakr, "AutoMLBench: A comprehensive experimental evaluation of automated machine learning frameworks," *Expert Systems with Applications*, vol. 243, p. 122877, 2024, Elsevier.

[7] R. Elshawi and S. Sakr, "Automated machine learning: Techniques and frameworks," in *Big Data Management and Analytics: 9th European Summer School, eBISS 2019, Berlin, Germany, June 30–July 5, 2019, Revised Selected Papers*, pp. 40–69, Springer, 2020.

[8] L. Liu, W. Zhou, K. Guan, B. Peng, S. Xu, J. Tang, Q. Zhu, J. Till, X. Jia, C. Jiang, *et al.*, "Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems," *Nature Communications*, vol. 15, no. 1, p. 357, 2024.

[9] J. Heil, J. M. Valencia, and A. Stein, "Towards crop yield prediction using Automated Machine Learning," 2023, Gesellschaft für Informatik eV.

[10] D. Prabha, R. Siva Subramanian, M. G. Dinesh, and P. Girija, "Sustainable Farming Through AI-Enabled Precision Agriculture," *Artificial Intelligence for Precision Agriculture*, p. 159, 2024, CRC Press.

[11] Cucurachi S, van der Giesen C, Guinée J (2018) Ex-ante LCA of emerging technologies. Procedia CIRP 69463–468

[12] van der Giesen C, Cucurachi S, Guinée J, Kramer GJ, Tukker A (2020) A critical view on the current application of LCA for new technologies and recommendations for improved practice. J Clean Prod 259120904

[13] Voglhuber-Slavinsky A, Zicari A, Smetana S et al (2022) Setting life cycle assessment (LCA) in a future-oriented context: the combination of qualitative scenarios and LCA in the agri-food sector. Eur J Futur Res 10:15

[14] Tsoy N, Prado V, Wypkema A, Quist J (2019) Anticipatory life cycle assessment of sol-gel derived anti-reflective coating for greenhouse glass. J Clean Prod 221:365–376

[15] https://www.suscrop.eu/funded-projects/3rd-call/connectfarms

[16] Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2020). Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074*, 24, 8.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[18] Hutter, F., Hoos, H. H., Leyton-Brown, K., & Stützle, T. (2011). Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization (LION 5)* (pp. 507-523). Springer.