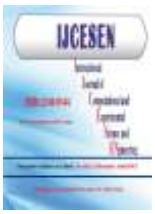
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 5799-5807
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Hybrid Optimization of Logistic Regression for Water Footprint Modeling in Iraqi Agriculture

Huda Mowafek Kadhim[1]*, Ali Hasan Taresh[2]

[1]Information Technology & Communications University Baghdad, Iraq
* **Corresponding Author Email:** ms202320754@iips.edu.iq- **ORCID:** 0000-0002-5247-1850

[2]Information Technology & Communications University Baghdad, Iraq
Email: alihtaresh@uoitc.edu.iq- **ORCID:** 0000-0002-8776-0322

Article Info:

DOI: 10.22399/ijcesen.3616
Received : 20 May 2025
Accepted : 30 July 2025

Keywords

Water footprint prediction, Logistic regression,
RFE,
SMOTE,
Feature scaling,
cross-validation

Abstract:

Water scarcity in Iraq underscores the urgent need for accurate water footprint (WF) prediction to support sustainable agricultural practices. This study presents an enhanced logistic regression (LR) model for WF forecasting by incorporating Recursive Feature Elimination (RFE), Synthetic Minority Oversampling Technique (SMOTE), and data normalization. RFE was employed to identify the most influential predictors, while SMOTE effectively addressed class imbalance within the dataset. Standard scaling was applied to stabilize model performance across varying data magnitudes. The model was evaluated using time-series cross-validation to ensure robustness and prevent data leakage, achieving a high predictive accuracy of 98.22%. The proposed framework offers a reliable tool for forecasting WF trends in Iraq over the period 2025–2030, contributing to evidence-based water resource management in arid agricultural regions.

1. Introduction

Al-Badri et al. 2023 They point to the escalation of water shortages in Iraq as a result of local demand and conflicts over shared rivers, especially in agriculture, which is considered the most water-consuming area, consuming more than 80% of freshwater resources, and therefore considered an important matter for achieving sustainability and food security [1]. Ewaid et al. 2019 They explain that due to the lack of rainfall, increased evaporation, high temperatures and inefficient irrigation systems, Iraq is facing increasing weakness in agriculture, and this affects the availability of water and crop production. They stress in the governorates of Iraq the necessity of measuring the water footprint [2].

De Mauro et al. 2022 explains that governments and businesses can make smarter decisions using a powerful tool, machine learning, by analyzing complex, big data sets. In agriculture, for example, it is used to monitor crop behavior, water usage, and future demand [3]. Isak-Zatega et al. 2020 shows that one of the most widely used algorithms is logistic regression for binary classification problems such as predicting whether crops are locally grown or imported. God makes it suitable for decision-making processes [4]. Katharria et al. 2025 focuses on the use of machine learning technology around smart agriculture in irrigation planning, crop monitoring, yield prediction, and disease detection, indicating an important role for precision analytics and data integration in agricultural innovation [5].Park and Kim 2020 explain that by selecting only the most important features, classification results are improved by using Recursive Feature Elimination (RFE). This makes the model more focused and simpler, which helps it generalize better to new data [6]. Hemmatian et al. 2025 proposed an improved version that helps balance a dataset by generating new samples for underrepresented classes, which is (SMOTE) [7]. Bhagat and Bakariya (2025) express to maintain the data's temporal order, the time-series cross-validation methods are used to make them suitable for evaluating models on future data that is unseen [8].

2. Related Works

In 2025, Emeç et al. established a global model using an ensemble machine learning approach with

AdaBoost to predict the water footprint of wheat, resulting in a total water footprint prediction accuracy of 97.35% and an R^2 of 69% [9].

In 2025, Mortazavizadeh et al. reviewed machine learning applications in agricultural water management, point out that models like ANN, Random Forest, and SVM noticeably improved estimation of water footprint and prediction of evapotranspiration, in some studies the Random Forest achieving up to 99% accuracy [10].

In 2024, Abdel-Hameed et al. The proposed system uses machine learning models like SVR, XGBoost, random forest and ANN to estimate the blue water footprint of potato crops, where ANN and XGBoost resulted in the highest accuracy of $R^2 > 0.95$ [11].

In 2024, Mabasha et al. designed a smart irrigation system using Decision Tree and SVM models, with SVM achieved an accuracy of 98.94%, which performed better than DT with an accuracy of 94.5% [12].

In 2024, Al-Taher et al. employed machine learning to predict the water footprint of sugarcane in Sudan using SVR, random forest, XGBoost and Hybris. The SVR model achieved the best $R^2 = 0.98$, and the prediction errors reduced hybrid models [13].

In 2023, Mahore and Gadge established a machine-learning model by using random forest, Naive Bayes and SVM for predicting crops, where random forest achieved the highest accuracy [14].

In 2022, Hina and Hasan presented a machine learning system using a Decision Tree algorithm for predicting crop yields across India, which achieved an accuracy of 95% [15].

In 2020, Patil et al. suggested a machine learning model using Decision Trees KNN and Linear Regression algorithms where for cotton, it achieved 99% with Linear Regression and for sugarcane, it achieved 98% [16].

In 2022, Sadri et al.developed a FarmCan model using random forest algorithm and remote sensing which is used to forecast the lack crop water. For needed irrigation achieved average R2 of 68% and for evapotranspiration prediction achieved KGE values of up to 0.71 across Canadian farms [17].

3. Problem Statement

Accurate forecasting of the agricultural water footprint is crucial for sustainable resource management, especially in water-stressed regions such as Iraq. This study proposes an optimized logistic regression model enhanced through a hybrid data preprocessing pipeline comprising the Synthetic Minority Oversampling Technique (SMOTE), Recursive Feature Elimination (RFE), and normalization. The dataset, derived from agricultural inputs and regional climatic parameters, exhibited significant class imbalance and multicollinearity challenges. The proposed approach addressed these issues by improving data dimensionality and quality before model training. The study demonstrated superior predictive performance in accuracy, precision, and recall compared to classical logistic regression. The results highlight the potential of integrated preprocessing techniques to enhance traditional machine learning algorithms for decision-making and environmental modeling in arid agricultural systems.

Mathematical Formulation:

Let $X \in Rn \times dX \in \mathb{R}^{n \times d}X \in Rn \times d$ represent the input matrix, where n is the number of samples and d is the number of features. The target vector $y \in \{0,1\}ny \in \{0,1\}^ny \in \{0,1\}n$ denotes whether the water footprint is low (0) or high (1). The objective is to train a logistic regression model f(X) that minimizes the classification loss. However, orignal models struggled with performance because of irrelevant features, data leakage and class imbalance.

To address these challenges:

- Recursive Feature Elimination (RFE) selects the optimal features: $X' = RFE(X)$
- SMOTE balances the class distribution: $(X'', y'') = SMOTE(X', y)$
- Standard Scaling normalizes the data: $X''' = StandardScaler(X'')$.

Finally, time-series cross-validation is used to avoid data leakage to guarantee robust predictions.

The model aims to optimize the logistic regression function f to maximize accuracy, recall, and precision.

4. Methodology

4.1 Model Limitations

These results point out a main issue. While the model presented a rather good recall (indicating its capacity to correctly identify most positive instances), the precision was low. This shows that the model made a large number of false positive predictions. The cases are actually negative, but the model classifies them as positive (high water footprint). The low precision weakens the reliability of the model, especially when it's vital to minimize

false positives in applications, like in the management of water resources.

Also, because of the use of random splits, the model performance was negatively affected in the training data, leading to a large variance in the results and data leakage. Randomly splitting data for testing and training without maintaining time series results in inconsistent evaluation of the model, keeping the model from effectively generalizing. These challenges point to the need for a more reliable and robust method for evaluating the model and also the need for enhanced techniques for selection and feature scaling. Dealing with issues of irrelevant features, data leakage and class imbalance to enhance the performance of the model and guarantee it can be used in scenarios like real-world agricultural water management.

4.2 Method of Solution

To improve the stability and performance of the logistic regression model, multiple improvements were performed. Firstly, to enhance feature selection, the Recursive Feature Elimination (RFE) was applied, making sure that the most relevant variables only were included in the model. Then StandardScaler was performed to normalize the data for feature scaling, dealing with potential differences in the magnitude of features. The Synthetic Minority Oversampling Technique (SMOTE) was applied to handle class imbalance, which creates synthetic instances for a class that is not well-represented, thus improving model learning. Also, to ensure the dataset balance, the resampling and retraining were performed, which is important for model convergence. To establish clear decision boundaries, the model's probability threshold was optimized and maintained at 0.5. An important change was switching from 30 random iterations to 10-fold time-series cross-validation, which prevented data leakage, improved the robustness of the model and preserved the time series of the data. Finally, for future predictions, consistent preprocessing was applied to guarantee that the model remains reliable. Two sources were used to assemble the dataset used for training and evaluation, which included environmental and crop-specific water usage attributes [18][19]. This enriched feature space contributed to improved model accuracy, enabling more reliable predictions for water footprint management in agriculture.

4.3 Model Evaluation Visuals

To support the evaluation of changes made to the logistic regression model, the following tables and charts were generated for analysis.

1. Initial Logistic Regression (30 Iterations)

Table 1 indicates important changes in model performance through the iterations. Iteration 1 showed only an accuracy of 0.7423, a recall of 0.7857 and a precision of 0.5000. The best was shown in iteration 5 with an accuracy of 0.9939, recall of 1.0000, and precision of 0.9778. Iterations 11 and 24 also performed strongly, with high precision and accuracy. Where iteration 27 shows the poorest results, with the lowest precision of 0.4878 and accuracy of 0.7055, iterations 4 and 6 showed high recall but low precision. To enhance the performance, multiple improvements were applied.Figure 1 is a line chart showing accuracy, recall, and precision across 30 iterations of the initial logistic regression model. Recalls stay high, while accuracy shows moderate variances. Precision varies significantly, with several sharp drops.

2. Improved Logistic Regression (30 Iterations)

Table 2 shows that all 30 iterations of the enhanced logistic regression model performed strongly. Perfect scores were achieved by iterations such as 5, 8, 15, and 30 (1.0000 of accuracy, recall, and precision). Even though iterations like 6, 10, and 13 showed the minimum performance, it still maintained high metrics with precision above 0.97 and accuracy above 0.98. It shows the model improvement after implementing SMOTE, RFE and normalization techniques.

3. Logistic Regression with Time-Series Cross-Validation

Table 3 shows that Folds 1 and 2 resulted in the lowest recall values, 0.8750 and 0.8846, but precision remained at 1.0000. where most folds show high accuracy above 0.97 and perfect precision. Which confirms the strong robustness of the model when evaluated in a timely manner.

4.4 Results

After Improvements (30-Iterations):

After applying these improvements, the model's performance improved significantly, resulting in an average accuracy of 0.9927, precision of 0.9920 and recall of 0.9934 throughout 30 randomised iterations, as shown in Figure 2.As shown in Figures 3 and 4, the performance of the improved logistic regression model after applying RFE, SMOTE, and scaling. In Figure 3, the line chart shows that accuracy, recall, and precision remain

consistently high across the 30 iterations, with minimal variances. Figure 4, the box plot, further confirms this stability and high medians for all metrics, indicating strong, reliable model performance with very few outliers.

Time-Series CV (10-Fold):

Then switch to 10-fold time-series cross-validation for realistic evaluation that shows robustness, which resulted in averages of 0.9822 accuracy, 0.9823 precision and 0.9536 recall. From the showing results, the model became more suited for agriculture applications like predicting crop water footprints, as shown in Figure 5.Figures 6 and 7 show the performance of logistic regression with time-series cross-validation. In Figure 6, the line chart shows high accuracy and precision across the 10 folds, while recall fluctuates more, especially in early and final folds. In Figure 7, the box plot supports this observation, with high medians and narrow spreads indicating stable performance and a wider range and lower median indicating variability in detecting true positives.

4.5 Comparison Performance and metrices

Table 4 compares the performance metrics of the improved logistic regression with decision tree [12] and random forest [11] classifiers using both 10-fold time-series cross-validation and 30-iterations. The proposed model surpasses both in all metrics, as seen in Table 4. It achieved the highest accuracy (0.9927), recall (0.9934), and precision (0.9920) in 30-iterations, and maintained better time-series CV with accuracy (0.9822), recall (0.9536), and precision (0.9823). These results show the proposed improvements, such as SMOTE, RFE, resampling and normalisation, effectively improve model predictive performance and robustness.

Table 1. *Performance metrics per iteration*

Iterations	Accuracy	Recall	Precision
1	0.7423	0.7857	0.5000
2	0.7607	0.8214	0.6133
3	0.8528	0.9556	0.6615
4	0.7669	0.9762	0.5256
5	0.9939	1.0000	0.9778
6	0.8221	1.0000	0.5915
7	0.9264	0.8431	0.9149
8	0.8098	0.9592	0.6184
9	0.8221	0.9348	0.6232
10	0.8221	0.9649	0.6707
11	0.9755	0.9231	0.9730
12	0.9080	0.9130	0.7925
13	0.9080	0.9423	0.8033
14	0.7607	0.9333	0.5385
15	0.7669	0.8491	0.6000
16	0.7975	0.9400	0.6104
17	0.7669	0.8824	0.5844
18	0.9448	0.7805	1.0000
19	0.8712	0.8400	0.7636
20	0.9509	0.8542	0.9762
21	0.7975	0.9388	0.6053
22	0.8834	0.9636	0.7571
23	0.7669	0.9524	0.5263
24	0.9632	0.8846	1.0000
25	0.8589	0.7609	0.7447
26	0.7423	0.9608	0.5506
27	0.7055	0.8696	0.4878
28	0.7791	0.7755	0.6032
29	0.8650	0.9615	0.7143
30	0.8650	0.8947	0.6538
AVG	0.8429	0.8892	0.7161

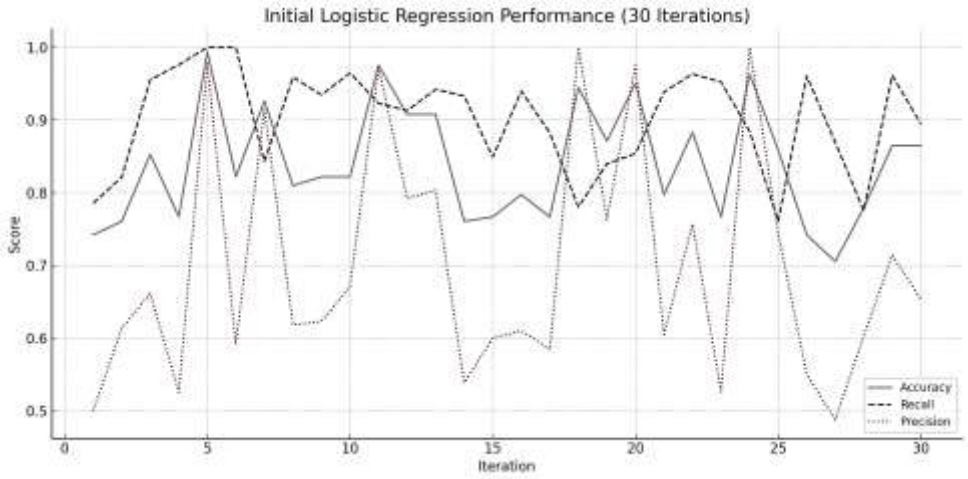


Figure 1. *Line chart: Accuracy, Recall, Precision over 30 iterations*

Table 2. *Metrics with RFE, SMOTE, and scaling*

Iterations	Accuracy	Recall	Precision
1	0.9914	0.9919	0.9919
2	0.9957	0.9910	1.0000
3	0.9914	0.9832	1.0000
4	0.9914	0.9832	1.0000
5	1.0000	1.0000	1.0000
6	0.9871	0.9831	0.9915
7	0.9914	0.9912	0.9912
8	1.0000	1.0000	1.0000
9	0.9871	0.9910	0.9821
10	0.9828	0.9917	0.9754
11	0.9871	1.0000	0.9754
12	0.9914	0.9915	0.9915
13	0.9871	0.9836	0.9917
14	0.9957	1.0000	0.9915
15	1.0000	1.0000	1.0000
16	0.9914	0.9916	0.9916
17	0.9828	0.9823	0.9823
18	0.9914	0.9909	0.9909
19	0.9914	0.9909	0.9909
20	0.9957	1.0000	0.9918
21	0.9957	1.0000	0.9921
22	0.9957	0.9915	1.0000
23	0.9914	0.9823	1.0000
24	0.9957	1.0000	0.9910
25	0.9957	1.0000	0.9918
26	0.9957	1.0000	0.9912
27	0.9914	0.9911	0.9911
28	0.9957	0.9914	1.0000
29	0.9957	1.0000	0.9905
30	0.9957	1.0000	0.9921

Table 3. *Fold-wise performance metrics*

Folds	Accuracy	Recall	Precision
1	0.9589	0.8750	1.0000
2	0.9589	0.8846	1.0000
3	1.0000	1.0000	1.0000
4	0.9863	0.9615	1.0000

5	0.9863	1.0000	0.9412
6	0.9863	1.0000	0.9444
7	0.9863	0.9444	1.0000
8	0.9726	0.9375	0.9375
9	1.0000	1.0000	1.0000
10	0.9863	0.9333	1.0000

```
Average Metrics Across Iterations:
Average Accuracy:   0.9927
Average Recall:     0.9934
Average Precision:  0.9920
```

Figure 2. *Metrics and Classification Report after 30-Iterations.*

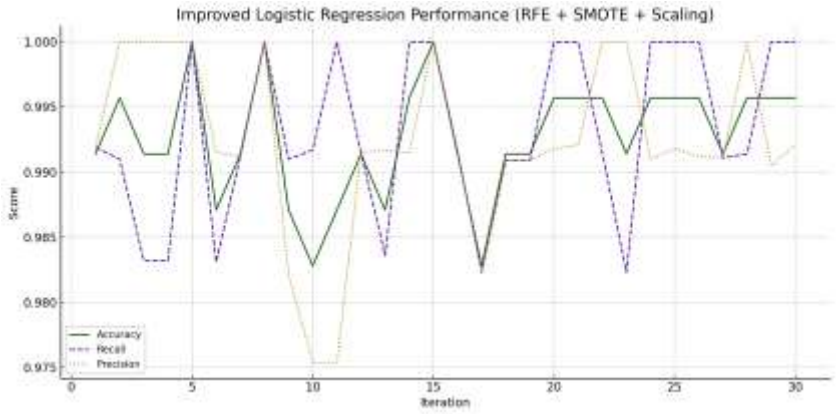


Figure 3. *Line chart: Improved iteration performance*

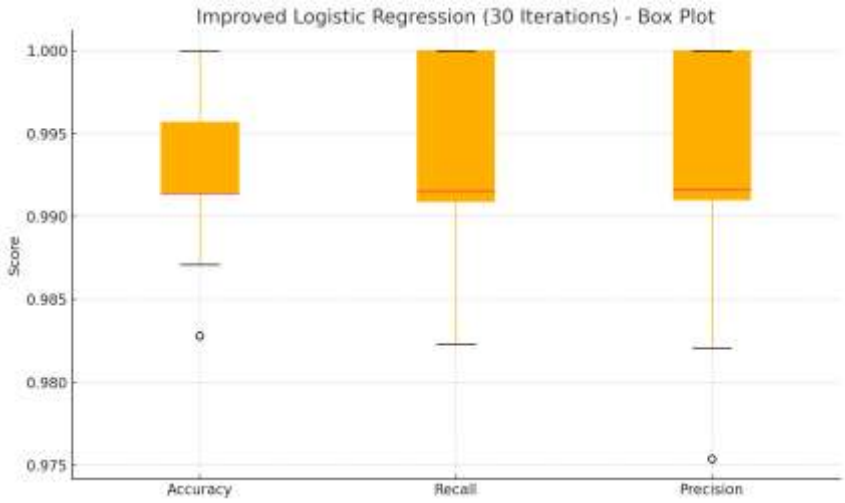


Figure 4. *Box plot: Improved metric distribution*

```
=============== Summary Statistics ===============
Metric          Average    Std Dev    Range
-----------------------------------------
Accuracy        0.9822     0.0138     0.9589-1.0000
Recall          0.9536     0.0451     0.8750-1.0000
Precision       0.9823     0.0271     0.9375-1.0000
```

Figure 5. *Metrics and Classification Report after 10-Fold.*

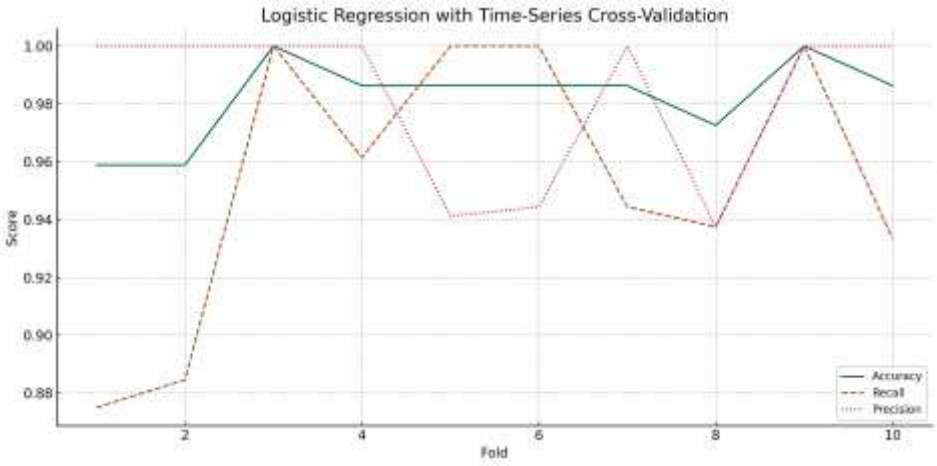


Figure 6. *Line chart across the 10 CV folds*

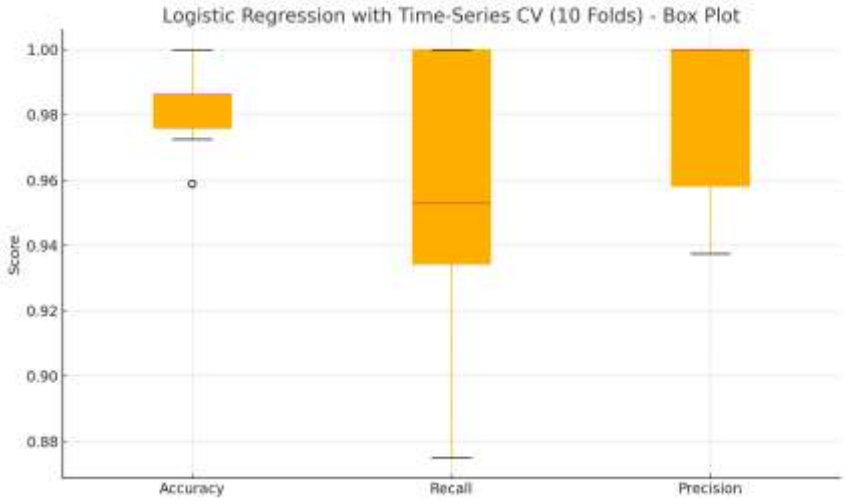


Figure 7. *Box plot: CV metric distribution*

Table 4. *Comparison Performance*

Method	How It Works	Issues	Advantages
30 Iterations	Randomly splits data 30 times (ignores time)	Risk of data leakage; overestimates performance	Fast, simple
Time-Series CV	Trains on past, tests on future (time-aware)	Requires time-sorted data	Realistic, prevents data leakage

	Algorithm / Metrices	Decision tree classifier [12]	Random forest classifier [11]	improved logistic regression (Proposed)
30-Iterations	Avg Accuracy	0.9855	0.9898	0.9927
	Avg Recall	0.9875	0.9908	0.9934
	Avg Precision	0.9837	0.9890	0.9920
Time-Series CV	Avg Accuracy	0.9575	0.9740	0.9822
	Avg Recall	0.9189	0.9385	0.9536
	Avg Precision	0.9390	0.9740	0.9823

4. Conclusions

For the prediction of water footprint in Iraq's agricultural sector, this study developed and enhanced a logistic regression model by integrating four key machine learning techniques: Recursive Feature Elimination (RFE) for feature selection, standardization for data normalization, SMOTE for addressing class imbalance, and resampling for improving data distribution. The optimized model

achieved an impressive accuracy of 98.22% through time-series cross-validation, with high recall (0.9536) and precision (0.9823). These results provided more reliable and realistic performance metrics, closely reflecting real-world scenarios. The model's exceptional precision makes it particularly valuable for water resource management, where the accurate identification of crops with high water footprints is critical for informed decision-making and sustainable agricultural practices. Future work could be aimed at enhancing prediction accuracy by adding more features and employing advanced ensemble methods. Building a model in real-world settings and using explainable AI would be helpful

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of 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
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] B. H. Al-Badri, M. K. Mohammad, and J. O. Khalid, "The Water Footprint and Virtual Water and Their Effect on Food Security in Iraq," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1222, no. 1, 2023, doi: 10.1088/1755-1315/1222/1/012023.

[2] S. H. Ewaid, S. A. Abed, and N. Al-Ansari, "Water footprint of wheat in Iraq," *Water (Switzerland)*, vol. 11, no. 3, pp. 1–12, 2019, doi: 10.3390/w11030535.

[3] A. De Mauro, A. Sestino, and A. Bacconi, "Machine learning and artificial intelligence use in marketing: a general taxonomy," *Ital. J. Mark.*, vol. 2022, no. 4, pp. 439–457, 2022, doi: 10.1007/s43039-022-00057-w.

[4] S. Isak-Zatega, A. Lipovac, and V. Lipovac, "Logistic regression based in-service assessment of mobile web browsing service quality acceptability," *Eurasip J. Wirel. Commun. Netw.*, vol. 2020, no. 1, 2020, doi: 10.1186/s13638-020-01708-2.

[5] K. D. Aashu Katharria, Kanchan Rajwar, Millie Pant, Juan D. Velásquez, Václav Snášel, "Information Fusion in Smart Agriculture: Machine Learning Applications and Future Research Directions," *arXiv Prepr. arXiv2405.17465*, 2025, doi: 10.48550/arXiv.2405.17465.

[6] T. Park and C. Kim, "Predicting the variables that determine university (Re-)entrance as a career development using support vector machines with recursive feature elimination: The case of South Korea," *Sustain.*, vol. 12, no. 18, 2020, doi: 10.3390/SU12187365.

[7] J. Hemmatian, R. Hajizadeh, and F. Nazari, "Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE," *PLoS One*, vol. 20, no. 2, p. e0317396, 2025, doi: 10.1371/journal.pone.0317396.

[8] M. Bhagat and B. Bakariya, "A Comprehensive Review of Cross-Validation Techniques in Machine Learning," vol. 16, no. 1, pp. 1–4, 2025.

[9] M. Emeç, A. Muratoğlu, and M. S. Demir, "High-resolution global modeling of wheat's water footprint using a machine learning ensemble approach," *Ecol. Process.*, vol. 14, no. 1, 2025, doi: 10.1186/s13717-025-00594-0.

[10] F. Mortazavizadeh *et al.*, "Advances in machine learning for agricultural water management: a review of techniques and applications," *J. Hydroinformatics*, vol. 27, no. 3, pp. 474–492, 2025, doi: 10.2166/hydro.2025.258.

[11] A. M. Abdel-Hameed *et al.*, "Estimation of Potato Water Footprint Using Machine Learning Algorithm Models in Arid Regions," *Potato Res.*, vol. 67, no. 4, pp. 1755–1774, 2024, doi: 10.1007/s11540-024-09716-1.

[12] Shaik Mabasha, "Evaluating the machine learning based Efficacy of Decision Tree and Support Vector Machines in Smart Irrigation Systems for Precise Irrigation Status Classification for Optimizing Water Management in Agriculture," *J. Electr. Syst.*, vol. 20, no. 2s, pp. 867–875, 2024, doi: 10.52783/jes.1682.

[13] R. H. Al-Taher *et al.*, "Predicting Green Water Footprint of Sugarcane Crop Using Multi-Source Data-Based and Hybrid Machine Learning Algorithms in White Nile State, Sudan," *Water (Switzerland)*, vol. 16, no. 22, 2024, doi: 10.3390/w16223241.

[14] Ruchira C. Mahore and Naresh G. Gadge, "Design A Model for Crop Prediction And Analysis Using Machine Learning," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 3307, pp. 90–95, 2023, doi: 10.32628/cseit228681.

[15] F. Hina and M. T. Hasan, "Agriculture Crop Yield Prediction Using Machine Learning," no. April, 2022.

[16] Ashwini I. Patil, R. A. Medar, and V. Desai, "Crop Yield Prediction Using Machine Learning Techniques," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 7, no. 3, pp. 312–315, 2020, doi: https://doi.org/10.32628/IJSRSET20736.

[17] S. Sadri, J. S. Famiglietti, M. Pan, H. E. Beck, A.

Berg, and E. F. Wood, "FarmCan: a physical, statistical, and machine learning model to forecast crop water deficit for farms," *Hydrol. Earth Syst. Sci.*, vol. 26, no. 20, pp. 5373–5390, 2022, doi: 10.5194/hess-26-5373-2022.

[18] O. Mialyk, J. F. Schyns, M. J. Booij, H. Su, R. J. Hogeboom, and M. Berger, "Water footprints and crop water use of 175 individual crops for 1990–2019 simulated with a global crop model," *Sci. Data*, vol. 11, no. 1, pp. 1–16, 2024, doi: 10.1038/s41597-024-03051-3.

[19] Google Earth Engine, "Google Earth Engine." Accessed: Oct. 15, 2024. [Online]. Available: https://earthengine.google.com/