International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Comparative Modeling of Cereal Yields in Senegal: A Hybrid Approach Using Bi-LSTM and Ensemble Learning on Agro climatic FAO Data

Pape El Hadji Abdoulaye Gueye[1*], Cherif Bachir Deme[2], Diery Ngom[3], Adrien Basse[4]
[1,2,3,4]Lecturer Researcher Department of TIC, UFR SATIC University Alioune Diop of Bambey, Bambey Senegal
Email address: [1]papeabdoulaye.gueye@uadb.edu.sn

Abstract*—Accurate prediction of cereal yields is critical for food security, particularly in Sahelian regions characterized by high climatic variability. This study develops a machine learning framework integrating dynamic agroclimatic variables (precipitation, temperature, soil nutrients) with FAO production statistics in Senegal over [2000–2024]. Feature selection based on correlation with yield indicated that MODIS-derived vegetation indices (NDVI, EVI, SAVI) were less relevant and thus excluded. Several models were evaluated, including Random Forest, XGBoost, CatBoost, and a Bidirectional LSTM explicitly designed to capture temporal dependencies. The Bi-LSTM achieved the highest predictive accuracy (R2 =0.94, RMSE=98.53), followed by CatBoost (R2=0.80,RMSE=216.21), XGBoost (R2=0.74, RMSE=243.07), and Random Forest (R2 = 0.72, RMSE = 251.46). Robustness was assessed using the Diebold-Mariano test, and interpretability was explored with SHAP values. The study demon strates that agroclimatic and production variables dominate over vegetation indices in predicting yields and highlights the trade-off between the superior accuracy of deep learning models and their higher computational cost. These results provide a reliable and interpretable framework for yield fore casting in Sahelian agriculture, emphasizing both methodological rigor and practical applicability*.

***Keywords*—** *Agricultural yields, Machine learn ing, Agroclimatic variables, Bi-LSTM, Senegal, Crop prediction*.

I. Introduction

Sahelian agriculture is particularly vulnerable to climatic fluctuations and environmental changes, making yield forecasting essential for anticipating production deficits, optimizing resource allocation, and strengthening the resilience of agricultural systems. However, developing accurate predictive models remains challenging due to the complex interactions between climate, soil, and agricultural practices.

Remote sensing products, particularly vegetation indices such as the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Soil-Adjusted Vegetation Index (SAVI), are widely used to monitor crop growth and estimate yields [1]. While these indices provide valuable insights into vegetation cover and biomass, their predictive power at the national scale is limited, due to vegetation saturation and the dominant role of climatic variability in shaping agricultural outcomes in the Sahel.

To overcome these limitations, the integration of FAO-reported production statistics with agroclimatic variables (precipitation, temperature, soil nutrient availability) is essential. Agricultural yield simulation models, such as DSSAT [2], WOFOST [3], PCSE [4], and APSIM [5], provide tools for predicting crop performance, but their application in Sub-Saharan Africa is constrained by limited local data, complex traditional cropping systems, and the need for advanced technical expertise.

Satellite imagery from MODIS and Sentinel-2 sensors allows tracking crop dynamics, yet models based solely on vegetation indices may lack sufficient resolution for small farms and often ignore local agroclimatic conditions [6–11]. The combination of climatic, soil, and remote sensing data with machine learning approaches (Random Forest, XGBoost, CatBoost) and recurrent neural networks such as LSTM has shown substantial potential to capture complex nonlinear relationships and improve yield prediction accuracy [14–16].

Despite these advances, key challenges remain, including data quality and availability, gaps in capturing local farming practices, and the generalization of models across heterogeneous regions. These limitations highlight the need for approaches tailored to local contexts, integrating agroclimatic, production, and remote sensing data with advanced computational methods for reliable yield prediction in Senegal and the wider Sahel.

The remainder of this paper is structured as follows: Section II describes the models and methods, including the dataset, variable selection, and predictive algorithms such as Random Forest, XGBoost, CatBoost, and Bidirectional LSTM. Section III presents the results and discussion, focusing on model comparison, feature importance, and interpretability using SHAP values. Section IV concludes the study by summarizing the key findings and contributions, while Section V outlines future perspectives and potential directions for further model enhancement.

II. Material and Methods

This study combines FAO agroclimatic data (precipitation, temperature, soil nutrients) with MODIS-derived vegetation indices (NDVI, EVI, SAVI) to model cereal yields in Senegal over multiple years.

To retain the most informative predictors, variables with an absolute correlation ≥ 0.5 with yield were selected, including cultivated area, fertilizer use (N, P, K), previous years' yields, and year of observation. The correlation matrix

Pape El Hadji Abdoulaye Gueye, Cherif Bachir Deme, Diery Ngom, and Adrien Basse, "Comparative Modeling of Cereal Yields in Senegal: A Hybrid Approach Using Bi-LSTM and Ensemble Learning on Agro climatic FAO Data," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 5, pp. 207-210, 2025.

in **Figure 1** shows these parameters. This improves model robustness and interpretability.

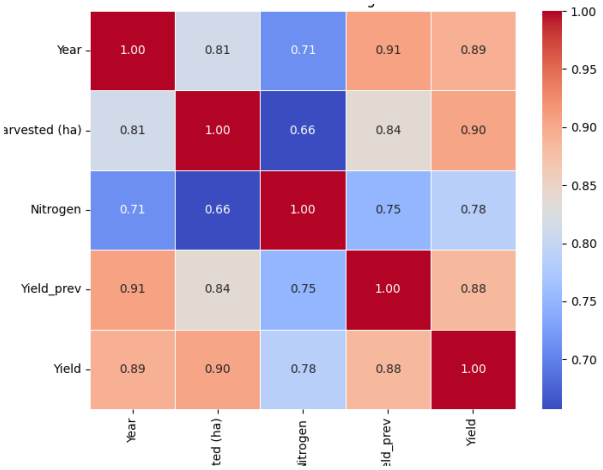

Figure 1: Correlation matrix after removing weakly correlated variables

Four predictive models were employed: Random Forest (RF), XGBoost, CatBoost, and Bidirectional LSTM (Bi-LSTM). Tree-based models efficiently capture nonlinear relationships in static data, while Bi-LSTM accounts for temporal dependencies in yield dynamics.

Model complexity:

RF is the fastest and least memory-intensive; boosting models (XGBoost, CatBoost) offer higher accuracy with moderate resources; Bi-LSTM is computationally demanding but best captures temporal trends. Empirical evaluation (Table 1) shows that Random Forest trains the fastest (0.05s) with minimal memory use (0.3MB), followed by XGBoost and CatBoost. Bi-LSTM, although slower (46.7s) and memory-intensive (317MB), captures temporal dependencies, which can improve accuracy.

TABLE 1: Training time and memory usage.

Model	Training Time (s)	Memory (MB)
Random Forest	0.05	0.30
XGBoost	0.04	3.45
CatBoost	0.24	7.50
Bi-LSTM	46.70	317.62

Hyperparameters:

Tree-based models used 100 estimators, depth 6, learning rate 0.1. Bi-LSTM employed 3 time steps, 3 bidirectional layers (280 units), separate branches for dynamic/static features, with Gaussian noise and early stopping for generalization. This framework enables reliable, interpretable prediction of cereal yields, integrating agroclimatic, soil, and temporal information. Overall, while Bi-LSTM is more complex, it is suited for modeling temporal agroclimatic trends. In contrast, tree-based models offer faster, resource efficient alternatives for static data scenarios.

III. Result and Discussion

A. Comparison of Predictive Models

Table 2 presents the performance of three machine learning models. CatBoost achieved the best results among tree-based methods, followed by XG Boost and Random Forest.

TABLE 2: Performance of ML models

Model	RMSE	R^2	Time (s)
Random Forest	251.46	0.72	0.05
XGBoost	243.07	0.74	0.04
CatBoost	216.21	0.80	0.24

Diebold-Mariano tests confirmed that CatBoost significantly outperformed RandomForest (DM = 4.306) and had a moderate advantage over XGBoost.

B. Bi-LSTM vs. CatBoost

The Bi-LSTM achieved the highest accuracy with an RMSE of 98.53 and R2 of 0.94 (Table 3). The Diebold-Mariano test (DM =-1.70) indicates a statistically significant advantage of Bi-LSTM over CatBoost. However, this comes at the cost of higher computational resources.

TABLE 3: Comparison after feature selection (r < 0.5 removed).

Model	RMSE	R^2
Bi-LSTM	98.53	0.94
CatBoost	216.21	0.80
XGBoost	243.07	0.74
Random Forest	251.46	0.72

C. Feature Importance and Model Interpretability

SHAP (SHapley Additive exPlanations) values were used to evaluate feature importance across models. The Bi-LSTM identified key agronomic drivers such as year, harvested area, nitrogen, and potassium, similarly to tree-based models. Figure 2 summarizes the SHAP outputs.

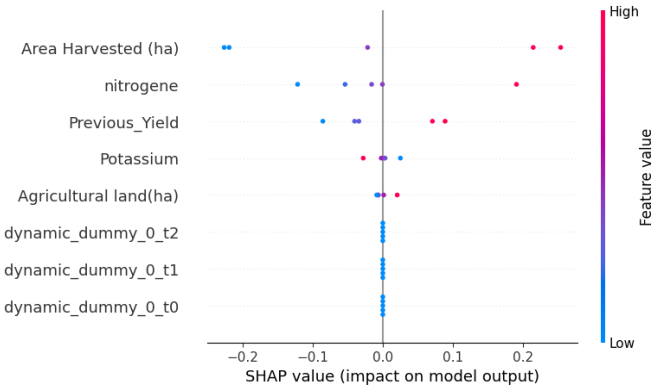

Figure 2: SHAP summary plot for Bi-LSTM model.

All models emphasized year, harvested area, previous yield, and nitrogen as key predictors. Bi-LSTM's temporal modeling provides more consistent attribution to past yields, enhancing its predictive power over tree-based models.

D. Discussion

The Bi-LSTM model achieved the highest accuracy, thanks to its ability to model temporal agro climatic patterns. SHAP analysis confirmed its reliance on key variables such as past yield, harvested area, and nitrogen use. Despite CatBoost

Pape El Hadji Abdoulaye Gueye, Cherif Bachir Deme, Diery Ngom, and Adrien Basse, "Comparative Modeling of Cereal Yields in Senegal: A Hybrid Approach Using Bi-LSTM and Ensemble Learning on Agro climatic FAO Data," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 5, pp. 207-210, 2025.

offering a solid balance between performance and efficiency, Bi-LSTM excelled in capturing time-dependent yield variability. However, this gain in precision comes at a computational cost, with significantly higher training time and memory usage. In low-resource settings, tree based models remain attractive alternatives. Over all, the results support the integration of temporal models like Bi-LSTM in forecasting frameworks, especially when seasonal dynamics are relevant. For operational use, hybrid models or cloud-based solutions could help reconcile accuracy with resource constraints.

IV. Conclusion

This study shows that the Bidirectional LSTM model offers the best performance for cereal yield prediction in Senegal, with an RMSE of 98.53 and an R2 of 0.94. Its ability to model temporal dependencies gives it an edge over traditional tree-based models. SHAP analysis also confirms its interpretability by highlighting the most influential agronomic variables. While tree-based models such as CatBoost, XG Boost, and Random Forest remain efficient and computationally lighter, they are less suited for cap turing time-dependent patterns.

V. Perspectives

Future work can enhance both accuracy and operational relevance through several directions:

• Model Optimization: Further hyperparameter tuning and architectural improvements could refine Bi-LSTM performance.

• Hybrid Approaches: Combining Bi-LSTM with tree-based models may yield better tradeoffs between accuracy and computation.

• Advanced Architectures: Exploring RNN variants, CNNs, or Transformer models may help capture complex spatiotemporal dynamics.

• Deployment: Techniques such as model compression and cloud-based inference can improve usability in real-time systems. • Generalization: Integrating additional data sources (e.g., satellite imagery, soil data) and testing across regions or seasons will improve robustness. These directions aim to enhance the robustness, scalability, and adaptability of crop yield prediction systems in diverse agro-climatic contexts.

Acknowledgement

The author sincerely acknowledges the financial and institutional support provided by Université Alioune Diop de Bambey (UADB), which made this research possible. Special thanks are also extended to Dr. Cherif Bachir Deme, Dr. Adrien Basse, and Dr. Diery Ngom for their valuable guidance, collaboration, and encouragement throughout this study.

Conflict Of Interest

The authors declare that there is no conflict of interest.

References

[1] C. J. Tucker, J. E. Vanpraet, M. J. Sharman, and G. Van Ittersum, “Satellite remote sensing of total dry-matter production in the Senegalese Sahel,” Remote Sensing of Environment, vol. 17, pp. 233–249, 1985.

[2] J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A. Hunt, and J. T. Ritchie, “The DSSAT cropping system model,” European Journal of Agronomy, vol. 18, no. 3–4, pp. 235–265, 2003.

[3] H. L. Boogaard, C. A. Van Diepen, R. P. Rotter, J. M. C. A. Cabrera, and H. H. Van Laar, “User’s guide for the WOFOST 7.1 crop growth simulation model and WOFOST Control Center 1.5,” Tech. Rep., SC-DLO, 1998.

[4] A. De Wit, H. Boogaard, D. Fumagalli, S. Janssen, R. Knapen, D. Van Kraalingen, and K. Van Diepen, “PCSE: Python Crop Simulation Environment,” Tech. Rep., Wa geningen UR, 2012.

[5] R. L. McCown, G. L. Hammer, J. N. G. Hargreaves, D. P. Holzworth, and D. M. Freebairn, “APSIM: a novel software system for model development, model testing and simulation in agricultural systems research,” Agri cultural Systems, vol. 50, no. 3, pp. 255–271, 1996.

[6] P. C. Doraiswamy, S. Moulin, P. W. Cook, and A. Stern, “Crop yield assessment from remote sensing,” Pho togrammetric Engineering and Remote Sensing, vol. 70, no. 6, pp. 687–692, 2004.

[7] B. Hall et al., “Agricultural systems,” 2018.

[8] O. Roupsard et al., “Remote sensing applications,” 2020.

[9] N. D. Mueller, P. C. West, M. Johnston, D. K. Ray, and N. Ramankutty, “Global vulnerability to food inse curity,” Environmental Research Letters*, vol. 7, no. 4, p. 045003, 2012.

[10] M. Roznik, M. Boyd, and L. Porth, “Improving crop yield estimation by applying higher resolution satellite NDVI imagery and high-resolution cropland masks,” Remote Sensing Applications: Society and Environment, vol. 25, p. 100693, 2022.

[11] L. Seguini, A. Vrieling, M. Meroni, and A. Nelson, “An nual winter crop distribution from MODIS NDVI time se ries to improve yield forecasts for Europe,” International Journal of Applied Earth Observation and Geoinformation, vol. 130, p. 103898, 2024.

[12] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[13] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[14] S. P. G. Tahi, V. R. Houndji, K. V. Salako, C. G. Hounmenou, and R. G. Kakaï, “Machine learning tech niques for cereal crops yield prediction: A comprehen sive review,” Applications of Modelling and Simulation, vol. 8, pp. 174–190, 2024. [Online]. Available: http://arqiipubl.com/ams.

[15] A. B. Sarr and B. Sultan, “Predicting crop yields in Sene gal using machine learning methods,” International Journal of Climatology, vol. 43, no. 4, pp. 1817–1838, 2023. [Online]. Available: https://doi.org/10.1002/joc.7947

[16] D. B. Lobell, G. Azzari, Z. Jiang, and S. Wang, “Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling,” Environmental Modelling & Software*, vol. 139, p. 104993, 2021.

[17] M. Schuster and K. K. Paliwal, “Bidirectional recur rent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[18] S. A. Shammi and Q. Meng, “Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling,” Ecological Indicators, vol. 121, p. 107124, 2021.

[19] B. Solly, J. Andrieu, E. H. B. Dieye, and A. M. Jarju, “Dy namiques contrastées de reverdissement et dégradation de la couverture végétale au Sénégal révélées par analyse de série temporelle du NDVI MODIS,” VertigO: La revue électronique en sciences de l’environnement, vol. 22, no. 1, 2022, doi: 10.4000/vertigo.35589.

[20] D. K. Ray, N. Ramankutty, N. D. Mueller, P. C. West, and J. A. Foley, "Yield trends are insufficient to double global crop production by 2050," PLoS ONE, vol. 8, no. 6, p. e66428, 2013.

[21] X. Zhu, C. Wang, Y. Chen, H. Tang, and L. Song, "Crop yield estimation using machine learning methods with remote sensing data," Remote Sensing, vol. 13, no. 5, p. 922, 2021.

Pape El Hadji Abdoulaye Gueye, Cherif Bachir Deme, Diery Ngom, and Adrien Basse, “Comparative Modeling of Cereal Yields in Senegal: A Hybrid Approach Using Bi-LSTM and Ensemble Learning on Agro climatic FAO Data,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 5, pp. 207-210, 2025.

[22] A. Diouf, F. Niang, and B. T. Ndiaye, "Utilisation du NDVI pour le suivi des rendements agricoles au Sénégal," Cahiers Agricultures, vol. 9, pp. 357–364, 2000.
[23] R. Fensholt and K. Rasmussen, "Analysis of trends in the Sahelian vegetation dynamics using GIMMS NDVI dataset (1981–2007)," Remote Sensing of Environment, vol. 115, no. 2, pp. 288–302, 2011.
[24] B. T. Ndiaye, A. Diouf, and S. T. Ba, "Modélisation de la production agricole au Sénégal à partir des données de télédétection," Revue Scientifique et Technique, vol. 18, pp. 45–58, 2019.
[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Fomenko, "CatBoost: Unbiased boosting with categorical features," arXiv preprint arXiv:1706.09516, 2017.
[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [27] E. Sarr et al., "Precision agriculture," 2018.

Pape El Hadji Abdoulaye Gueye, Cherif Bachir Deme, Diery Ngom, and Adrien Basse, "Comparative Modeling of Cereal Yields in Senegal: A Hybrid Approach Using Bi-LSTM and Ensemble Learning on Agro climatic FAO Data," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 5, pp. 207-210, 2025.