AI-based algorithm for the management and optimization of smart agricultural IoT system

Aya Saad
LASEE Laboratory
University of Sousse
Higher Institute of
Computer Science and
Communication Technologies
4011, Sousse, Tunisia
Email: saadaya13@gmail.com

Ferdaws Ben Naceur
University of Sousse,
National Engineering
School of Sousse,
Laboratory of Advanced
Technology and Intelligent
Systems LATIS,
4023, Sousse, Tunisia
Email: swadref87@hotmail.com

Chokri Ben Salah
LASEE Laboratory,
Department of Electrical Engineering,
University of Monastir 5035;
ISSAT of Sousse, 4003 Tunisia
Email: chokribs@yahoo.fr

***Abstract*—Efficient management of agricultural water resources has become increasingly critical due to climate variability and rising global food demand. This paper presents a comprehensive IoT-based system for real-time agricultural water forecasting, integrating field-deployed sensors, cloud infrastructure, and advanced machine learning models. The system automates data collection, preprocessing, and model training, enabling accurate and scalable irrigation management. We evaluate three models: a lightweight XGBoost regressor for edge deployment, a Long Short-Term Memory (LSTM) network for capturing temporal patterns, and a hybrid LSTM–XGBoost model that combines the strengths of both. The hybrid model achieved the best performance with a Root Mean Squared Error (RMSE) of 0.01705 and a coefficient of determination (R^2) of 0.95, outperforming the standalone XGBoost (RMSE = 0.0184, R^2 = 0.92) and LSTM (RMSE = 0.0704, R^2 = 0.86) models. Operational insights regarding system latency, data reliability, and field maintenance are also discussed, emphasizing the model's robustness and practical deployment potential. The results underscore the viability of data-driven irrigation forecasting for improving agricultural sustainability and optimizing resource efficiency.**

***Index Terms*—IoT, LoRaWAN, Precision Agriculture, LSTM, XGBoost, Agricultural Water Forecasting**

I. Introduction

Agriculture consumes nearly 70% of global freshwater withdrawals, and climate induced variability in precipitation and temperature further stresses water availability. This underscores the urgent need for adaptive irrigation strategies to ensure food security and sustainability [1]. Traditional irrigation, based on fixed schedules or manual assessments, often leads to inefficient water use and ecological consequences [3]. Recent advances in machine learning have shown promise in optimizing these practices by improving water use forecasting and management [2].

In parallel, the Internet of Things (IoT) has enabled real-time environmental sensing through low-power networks like LoRaWAN [4]. However, integrating this continuous data flow with intelligent forecasting remains challenging. Existing solutions often emphasize either edge analytics [5] or centralized cloud pipelines [6], without offering unified, scalable systems that bridge both domains.

Moreover, while deep learning models can capture complex relationships, their high computational demands limit deployment on resource-constrained farms [7]. In contrast, lightweight models combined with efficient preprocessing can deliver accurate predictions at scale. Yet most prior systems still focus on individual components either the sensing, communication, or modeling layer without offering a cohesive, deployable framework.

To address these gaps, we propose a unified IoT machine learning architecture tailored to medium scale agricultural environments. Our system features cloud-based training and deployment of a hybrid LSTM-XGBoost model that leverages the sequential learning capacity of LSTMs and the predictive robustness of XGBoost. This configuration delivers high accuracy, centralizes model management, and remains adaptable to diverse field conditions.

The key contributions of this work are:

1) **Modular IoT System**: Design and deployment of a robust IoT infrastructure for real time irrigation forecasting, integrating in field sensors with a secure communication and cloud-based inference pipeline.
2) **Hybrid Forecasting Model**: Development of a hybrid LSTM-XGBoost model that combines temporal pattern extraction and gradient-boosted regression for accurate water efficiency prediction.
3) **Benchmarking and Validation**: Comprehensive performance evaluation against recent IoT enabled forecasting studies, demonstrating improved accuracy and operational feasibility.

The remainder of this paper is organized as follows: Section II reviews related work; Section III details system design and methods; Section IV outlines data and preprocessing; Section V presents the modeling approach; Section VI offers discussion; Section VII evaluates results; and Section VIII concludes with future directions.

II. Related Work

IoT integration in precision agriculture has progressed rapidly, notably through wireless sensor networks such as LoRaWAN [8] and energy efficient communication protocols [22]. These technologies support continuous, low-power monitoring of environmental conditions across large-scale farms. Complementary advances in cloud computing have enabled scalable data aggregation and remote decision making.

To reduce latency and offload computation, edge based frameworks like EdgeLSTM [5] process sensor data locally, improving responsiveness in real time irrigation control. Yet, full integration of edge analytics, cloud infrastructure, and predictive modeling remains uncommon.

On the modeling side, recent surveys have evaluated SVR, RF, CNNs, and XGBoost in agricultural forecasting [12], [18]. XGBoost is often favored for its predictive accuracy and computational efficiency. Hybrid models such as attention enhanced LSTM with boosting have emerged to better capture nonlinear temporal dependencies while retaining interpretability [17].

Past efforts like Gill et al. [19] and Ravi et al. [21] applied SVR and XGBoost to soil and crop forecasting, while Hossain et al. [20] explored image-based methods. However, these typically assume stable connectivity and do not address practical deployment issues such as communication constraints, sensor reliability, or edge cloud orchestration.

Federated learning has been proposed to mitigate data privacy concerns in distributed systems [7], yet its adoption in precision irrigation remains limited. Still, it offers a scalable path forward, especially in multi farm scenarios with heterogeneous infrastructure.

Unlike these fragmented approaches, our work delivers a unified architecture combining real time IoT sensing, secure cloud communication, and a hybrid LSTM-XGBoost model. This setup enables centralized deployment while leveraging LSTM's sequence learning and XGBoost's robustness for regression—bridging the gap between theoretical modeling and field-ready forecasting.

Overall, while individual components wireless sensing, edge computing, and ML models are well studied, holistic end-to-end systems that integrate all layers into a practical, deployable pipeline are scarce. This paper contributes to closing that gap.

III. Materials and Methods

We propose a modular, five layer IoT architecture tailored for real-world agricultural deployments. These layers Sensor Layer, Edge Gateway, Communication Stack, Cloud Backend, and Operational Controls collect, transmit, and process environmental data for intelligent irrigation forecasting. The architecture emphasizes robustness, low latency, and scalability.

A. Sensor Layer

The system integrates environmental and operational sensors:

- **Soil Moisture/Temperature Sensors:** Monitor root zone water levels and thermal conditions.

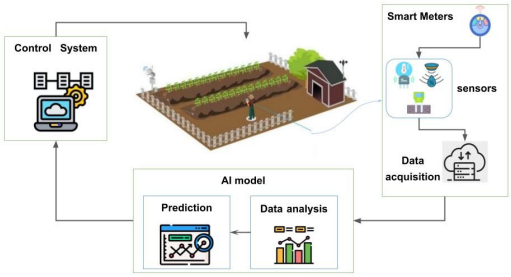


Fig. 1: Overview of the proposed smart irrigation forecasting system.

- **Ambient Climate Modules:** Capture temperature and humidity affecting evapotranspiration.
- **Water Flow Meters:** Measure irrigation volume and track usage patterns.
- **Pump Energy Monitors:** Log energy consumption to detect inefficiencies.

B. Edge Gateway and Inference

Edge devices (e.g., Raspberry Pi 4) collect data via LoRaWAN, apply timestamp alignment and filtering, and run lightweight **XGBoost** models for rapid, local inference. This ensures responsiveness under limited connectivity.

The cloud hosts the more complex **Hybrid LSTM-XGBoost model**. The LSTM captures temporal patterns; its latent representations are passed to XGBoost, which performs the final regression. This separation balances edge autonomy with cloud level precision.

C. Communication Protocols

A dual-layer stack supports reliable, secure data flow:

- **LoRaWAN:** Long-range, low-power transmission between sensors and gateways.
- **MQTT over TLS:** Lightweight, encrypted message delivery from gateway to cloud.

This setup minimizes energy use and ensures secure operation across remote sites.

D. Cloud Backend

Cloud processing is managed via **AWS IoT Core** and **AWS Lambda**. Incoming data undergoes:

- Mean imputation for missing values,
- IQR-based outlier filtering,
- One-hot encoding of categorical features,
- Normalization of continuous variables.

Processed data feeds model training pipelines and updates. The LSTM encodes complex dependencies into feature embeddings, which XGBoost refines to predict irrigation efficiency. Deploying this hybrid model in the cloud avoids edge limitations while supporting centralized updates and system scalability.

IV. Data Description and Preprocessing

A. Data Description

The dataset was constructed by merging two sources: a Crop Recommendation Dataset (with features such as soil nutrients, moisture, climate, pH, and irrigation method) and an Agricultural Water Usage Dataset (with water consumption and allocation records). They were integrated via crop type and region, yielding 2,200 complete samples covering multiple cycles and zones. The target variable, *irrigation efficiency*, represents the ratio of water consumed to allocated water per hectare a key sustainability indicator. The dataset is balanced across efficiency levels, supporting robust supervised learning.

B. Data Preprocessing

We applied a multi-stage preprocessing pipeline to ensure data quality, interpretability, and modeling stability.

Target Variable Engineering: Irrigation efficiency was calculated as the ratio of consumed to allocated water. As shown in Figure 2, the distribution is unimodal and centered near 1.0, indicating mostly balanced irrigation behavior.

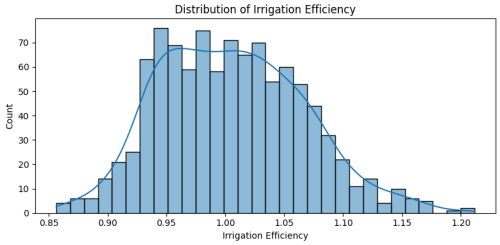
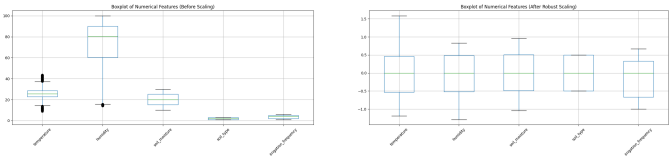


Fig. 2: Distribution of computed irrigation efficiency.

Feature Normalization: Numerical variables (e.g., temperature, humidity, soil moisture) were scaled using the RobustScaler to reduce the influence of outliers. Figure 3 contrasts distributions before and after scaling.

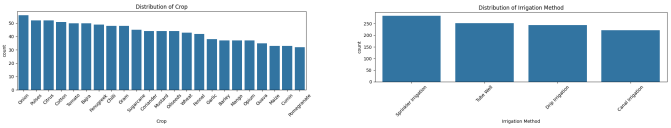


(a) Before (b) After

Fig. 3: Boxplots of numerical features before and after robust scaling.

Categorical Encoding: Features such as *Crop*, *District*, and *Irrigation Method* were encoded using one hot encoding to preserve class distinctiveness. Figure 4 shows example distributions for crop type and irrigation methods.

Feature Analysis: To explore feature dependencies, we generated a correlation matrix (Figure 5). Moderate correlations highlight the need for non linear models capable of capturing complex interactions.



(a) Crop Types (b) Irrigation Methods

Fig. 4: Distributions of selected categorical features.

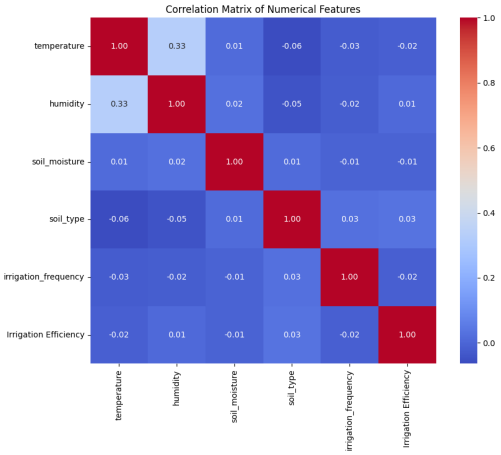


Fig. 5: Correlation heatmap for numerical features and irrigation efficiency.

V. Model Description

This section details the regression models employed for forecasting: LSTM, XGBoost Regressor and Hybrid Model (LSTM + XGBoost).

A. LSTM Model

As part of the modeling pipeline, a Long Short Term Memory (LSTM) network was implemented to explore deep learning's ability to capture complex relationships among environmental, agronomic, and irrigation features. Although the dataset was non temporal, inputs were reformatted into pseudo-sequences to enable compatibility with recurrent architectures.

The LSTM model was trained using standard supervised learning procedures, with features preprocessed through one hot encoding, normalization, and imputation. During training, the model learned latent representations that capture high-level interdependencies between inputs, such as the interaction between soil moisture, temperature, humidity, and irrigation strategy.

This model was evaluated using a held out validation set. It achieved a root mean squared error (RMSE) of **0.07044** and an R^2 score of **0.860**, demonstrating that LSTM networks can effectively model irrigation water consumption in structured sensor-driven agricultural datasets even in the absence of explicit temporal dynamics.

B. XGBoost Model

The XGBoost (Extreme Gradient Boosting) model is a tree-based ensemble learning algorithm designed to optimize

prediction accuracy through gradient boosting. Its architecture relies on sequential decision trees, each correcting the residuals of its predecessor. The model supports built-in regularization, handles missing values natively, and is optimized for performance and scalability.

In this work, XGBoost was trained using a pipeline that included mean imputation for numerical features and one-hot encoding for categorical attributes such as district, crop type, and irrigation method. A randomized grid search over hyperparameters (*n_estimators*, *learning_rate*, and *max_depth*) was conducted using a predefined validation split. The objective was to minimize mean squared error while maximizing R^2 on unseen data.

The model achieved a root mean squared error (RMSE) of 0.0184 and an R^2 score of 0.92 on the test set. Due to its robustness, low-latency inference, and low computational footprint, the XGBoost model was deployed on edge devices for real time water consumption forecasting in field conditions.

C. Hybrid LSTM-XGBoost Model

To exploit the complementary strengths of deep neural networks and gradient boosting algorithms, we designed a hybrid model that integrates Long Short Term Memory (LSTM) networks with XGBoost regression. The purpose of this architecture is to combine the feature learning capabilities of LSTM with the robustness and interpretability of tree-based ensembles.

The pipeline operates in two stages. First, an LSTM model is trained on the full input feature set, reformatted as pseudo-sequences to accommodate the network's structure. Rather than relying on raw predictions, the intermediate latent representations generated by the LSTM are extracted. These representations, optionally concatenated with selected input features, are then passed to a downstream XGBoost regressor to perform the final prediction.

This architecture allows the LSTM to capture nonlinear interdependencies and higher level abstractions in the data, which are subsequently refined by XGBoost's gradient-boosted decision trees. The hybrid model demonstrated improved generalization compared to both standalone LSTM and XGBoost implementations.

Empirically, the hybrid model achieved a root mean squared error (RMSE) of **0.01705** and an R^2 score of **0.950** on the validation set, outperforming all baseline models. This confirms the effectiveness of hybridization for modeling structured but complex agricultural datasets.

D. Model Complementarity

The proposed pipeline combines models with distinct strengths. LSTM networks are well suited for capturing complex, non linear relationships across environmental and irrigation parameters, while XGBoost is effective at modeling structured feature interactions and handling noise in heterogeneous tabular data. Their integration in a hybrid architecture allows the system to benefit from both learned latent representations and robust gradient boosted refinement. This complementarity significantly improved forecasting accuracy compared to standalone models.

VI. Experimental Results

We evaluated the proposed models—**XGBoost**, **LSTM**, and the hybrid **LSTM-XGBoost** on a held-out 20% test set. The goal was to assess forecasting accuracy, generalization, and operational trade offs.

A. Test-Set Performance

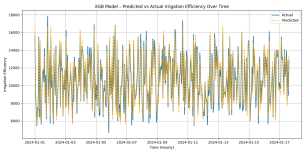
Table I presents Root Mean Squared Error (RMSE) and R^2 scores. XGBoost demonstrated strong accuracy with minimal overhead, while LSTM performed moderately despite the non temporal input format. The hybrid model outperformed both, confirming its advantage in combining sequence learning and tree based regression.

TABLE I: Performance of Forecasting Models on the Test Set

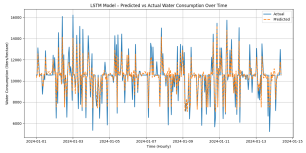
Model	**RMSE**	R^2
XGBoost	0.0184	0.92
LSTM	0.0704	0.8604
Hybrid LSTM–XGBoost	**0.01705**	**0.95**

B. Prediction Accuracy Visualization

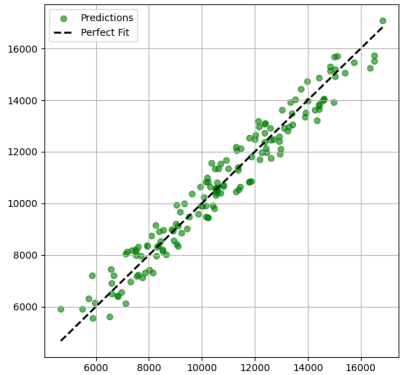
Figure 6 summarizes model performance. While both XGBoost and LSTM track observed values over time, the hybrid model yields tighter alignment with actual consumption, as reflected in its scatter plot.



(a) XGBoost



(b) LSTM



(c) Hybrid Model (Scatter)

Fig. 6: Predicted vs. Actual Water Consumption Across Models

C. Comparison with Prior Studies

Table II compares our results with recent IoT-based irrigation forecasting models. The hybrid model yields the lowest RMSE, demonstrating superior generalization over traditional SVR and prior XGBoost implementations.

TABLE II: Comparison with Recent IoT-Based Forecasting Studies

Study	Model	RMSE
Gill et al. (2006)	SVR	0.1392
Riaz Hossain et al. (2023)	SVR	0.0600
Ravi et al. (2020)	XGBoost	0.0240
Our Work (XGBoost)	XGBoost	0.0184
Our Work (LSTM)	LSTM	0.0704
Our Work (Hybrid)	LSTM–XGBoost	**0.01705**

D. Operational Insights

- **XGBoost** offers high accuracy with low latency, ideal for edge-based deployments.
- **LSTM** captures non-linear interactions but is more prone to noise and drift.
- **Hybrid** achieves the best precision and is optimal for centralized, high-resolution forecasting.

VII. Discussion

Beyond predictive performance, several practical considerations emerged during the design and deployment of the proposed system.

First, the decision to deploy the hybrid LSTM-XGBoost model entirely in the cloud was driven by computational requirements. The LSTM component, although powerful in capturing complex interactions, is not suitable for edge devices due to its resource demands. Offloading inference to the cloud ensures accuracy, but introduces latency that may not be tolerable in ultra low latency scenarios. However, in our use case daily or periodic irrigation planning rather than second by second actuation this trade off is acceptable.

Second, the system's reliance on stable communication links (LoRaWAN and MQTT over TLS) introduces points of vulnerability in rural deployments. While LoRaWAN provides long range coverage, factors like packet collision, signal attenuation, and gateway failures can lead to data loss or delays. Ensuring redundancy through multi gateway setups or caching at the gateway level is essential for robustness.

Third, environmental interference and sensor degradation are non trivial risks. Soil moisture sensors, for instance, can exhibit drift or failure over time. To address this, we employ robust preprocessing (IQR filtering, imputation) and recommend periodic calibration of sensors in long-term deployments.

Finally, while the system currently operates as a centralized cloud-hosted service, future extensions could explore federated learning or edge augmented inference to improve resilience, data privacy, and autonomy under intermittent connectivity conditions. These strategies would also help decentralize control and make the system more scalable across multiple farms or disconnected zones.

Security and Privacy Considerations

To ensure the integrity and reliability of field-deployed IoT systems, several security and privacy concerns must be addressed. Our current architecture leverages MQTT over TLS for secure, encrypted communication between edge gateways and the cloud, which protects against basic interception and man in the middle attacks. However, additional risks such as sensor spoofing, replay attacks, or false data injection remain plausible, particularly in unmonitored rural environments. To mitigate these risks, future implementations may incorporate lightweight authentication protocols at the sensor level, cryptographic device IDs, and anomaly detection algorithms capable of flagging suspicious data patterns. These measures would enhance trust in the decision-making pipeline and support secure, long-term scalability across distributed agricultural networks.

VIII. Conclusion and Future Work

This study presented a complete IoT-enabled machine learning framework for real-time forecasting of agricultural water consumption. The proposed system integrates in-field environmental sensing with a LoRaWAN–MQTT communication layer, a cloud based data processing pipeline, and a hybrid machine learning architecture that combines Long Short-Term Memory (LSTM) networks with XGBoost regression. This hybrid model is designed to capture both complex temporal dependencies and structured feature interactions in environmental and agronomic data.

Experimental results demonstrated that the hybrid LSTM-XGBoost model achieved the best performance (RMSE = 0.01705, R^2 = 0.95), outperforming both the standalone XGBoost model (RMSE = 0.0184, R^2 = 0.92) and the LSTM model (RMSE = 0.0704, R^2 = 0.8604). These findings validate the effectiveness of the hybrid design in capturing both sequential and structural data dependencies, and its superiority over recent IoT-based forecasting benchmarks in the literature. The system also highlights the practicality of centralized inference for complex models in resource-constrained agricultural settings.

Future Work

To further enhance the system's scalability and real-world applicability, several research directions are proposed:

- **Federated Learning Across Farms:** Implement decentralized learning strategies that allow individual farms to collaboratively train global models without sharing raw data. This supports privacy preservation and enables adaptation across heterogeneous agricultural environments.
- **Secure Communication Protocols:** Extend the current MQTT over TLS stack with lightweight cryptographic authentication schemes and anomaly detection algorithms to prevent spoofed sensor data or replay attacks.
- **Edge-Enabled Hybrid Inference:** Investigate low-power inference accelerators (e.g., Coral TPU, NVIDIA Jetson Nano) to evaluate the feasibility of running portions of the hybrid model locally under limited connectivity.
- **Multi-Agent Water Allocation Optimization:** Explore the use of multi-agent reinforcement learning (MARL) to model water distribution decisions between farms or

zones, aiming to optimize global water use under shared constraints.

- **Cross Climate Model Transferability:** Evaluate the generalizability of the current hybrid model in diverse climatic, soil, and irrigation contexts using transfer learning or meta-learning techniques.

Through these extensions, this work lays the foundation for next-generation precision agriculture solutions that combine AI, IoT, and cloud intelligence to support scalable, sustainable, and data-driven irrigation practices.

REFERENCES

[1] A. G. Koutroulis, L. V. Papadimitriou, M. G. Grillakis, I. K. Tsanis, K. Wyser, and R. A. Betts, "Freshwater vulnerability under high end climate change. A pan-European assessment," *Science of The Total Environment*, vol. 613, pp. 271–286, 2018. DOI: 10.1016/j.scitotenv.2017.09.197.

[2] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018. DOI: 10.1016/j.compag.2018.02.016.

[3] H. G. Jones, "Irrigation scheduling: advantages and pitfalls of plant-based methods," *Agricultural Water Management*, vol. 204, 2018.

[4] Y. Zhang and X. Liu, "A LoRaWAN-based IoT sensing system for smart agriculture," *IEEE IoT Journal*, vol. 7, no. 5, 2020.

[5] L. Cai, P. Wang, and J. Zhang, "EdgeLSTM: Real-time forecasting on edge devices," in *Proc. IoTDI*, 2021.

[6] J. Hernandez and T. Pilgrim, "Serverless IoT processing for agriculture," *Computers and Electronics in Agriculture*, vol. 190, 2021.

[7] M. Silva and R. Alves, "Federated learning for smart agriculture," *IEEE Transactions on Sustainable Computing*, vol. 7, 2022.

[8] J. Novak and E. Perez, "Evaluating LoRaWAN coverage for agriculture," *IEEE Transactions on Wireless Communications*, vol. 22, 2023.

[9] A. Rodriguez and J. Gutierrez, "IoT data preprocessing techniques," *Future Generation Computer Systems*, vol. 108, 2020.

[10] M. Smith and J. Clark, "Outlier-resistant preprocessing," *IEEE Sensors Journal*, vol. 20, 2020.

[11] A. Patil and D. Singh, "SVR for real-time soil moisture forecasting," *IEEE Transactions on Industrial Informatics*, vol. 19, 2023.

[12] P. Kumar et al., "XGBoost for crop water prediction," *Computers and Electronics in Agriculture*, vol. 197, 2022.

[13] C. Li and W. Zhao, "Stacking methods for irrigation forecasting," *Computers and Electronics in Agriculture*, vol. 189, 2021.

[14] D. Lopez and M. Ravinder, "Packet loss in LoRaWAN networks," *IEEE IoT Journal*, vol. 9, 2022.

[15] B. Sun and H. Zheng, "Real-time soil moisture forecasting," *IEEE Sensors Journal*, vol. 23, 2023.

[16] M. Acharya et al., "ML for hydrological forecasting," *Journal of Hydrologic Engineering*, vol. 26, 2021.

[17] H. Tang and Z.-L. Li, "Attention-based LSTM for crop prediction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, 2022.

[18] L. Zhang and Y. Wei, "Chlorophyll estimation using XGBoost and DL," *Plants*, vol. 12, 2023.

[19] M. K. Gill, T. Asefa, M. W. Kemblowski, and M. McKee, "Soil moisture prediction using support vector machines," *Journal of the American Water Resources Association*, vol. 42, no. 4, pp. 1033–1046, 2006. DOI: https://doi.org/10.1111/j.1752-1688.2006.tb04512.x.

[20] M. R. H. Hossain and M. A. Kabir, "Machine Learning Techniques for Estimating Soil Moisture from Mobile Captured Images," *arXiv preprint arXiv:2303.11527*, 2023. Available at: https://arxiv.org/abs/2303.11527.

[21] R. Ravi and R. Venkatesan, "Crop Yield Prediction using XG Boost Algorithm," *International Journal of Recent Technology and Engineering*, vol. 8, no. 5, pp. 3516–3520, 2020. DOI: https://doi.org/10.35940/ijrte.D9547.018520.

[22] L. García, L. Parra, J. M. Jiménez, and J. Lloret, "IoT-Based Smart Irrigation Systems: An Overview on the Recent Trends on Sensors and IoT Systems for Irrigation in Precision Agriculture," *Sensors*, vol. 20, no. 4, p. 1042, 2020. DOI: 10.3390/s20041042.