IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

AIoT-Based Precision Agriculture System Using Multi-Sensor Data Fusion and Predictive Analytics for Smart Irrigation and Crop Management

[1]Irfan Ahmad, [2]Md Imran Ansari
[1]M.Tech Scholar, [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]Patel College of Science and Technology, Bhopal, Madhya Pradesh, India

Abstract: Agriculture is increasingly adopting intelligent technologies to improve productivity and optimize resource utilization. This paper presents a simulation-based AIoT (Artificial Intelligence of Things) framework for precision agriculture that integrates multi-sensor data fusion and predictive analytics to enable smart irrigation and proactive crop management. Unlike conventional IoT systems that depend on physical sensor deployment, the proposed approach utilizes publicly available real-world agricultural datasets — including NASA SMAP soil moisture, NASA POWER meteorological data, and USDA SCAN field measurements — to create a reproducible, cost-free simulation of a six-sensor IoT-enabled farm environment. The system performs a rigorous data-preprocessing and feature-fusion pipeline, constructing a 47-feature matrix that captures temporal dynamics, crop-stress indicators, and multi-sensor interactions. Six machine learning algorithms — Linear Regression, Decision Tree, Random Forest, XGBoost, Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) — are benchmarked under identical experimental conditions. The LSTM model achieves state-of-the-art prediction accuracy of 96.8% ($R^2 = 0.97$, RMSE = 0.020 cm^3/cm^3 at 1-hour horizon), while XGBoost offers the best efficiency-accuracy trade-off for edge deployment ($R^2 = 0.96$, training time = 9.8 s). The intelligent irrigation decision-support engine generates crop-specific, volume-precise irrigation schedules that reduce water consumption by 42.3% relative to fixed-interval irrigation and by 23.8% relative to threshold-based reactive systems. Estimated crop yield improves by 12–18% by maintaining optimal soil moisture for 87.3% of the growing season. A full cloud-native, three-tier (Edge–Fog–Cloud) deployment blueprint with Docker/Kubernetes microservices ensures the simulation is directly translatable to physical IoT infrastructure with sub-300 ms end-to-end latency.

Index Terms — AIoT, Precision Agriculture, Multi-Sensor Data Fusion, Predictive Analytics, Smart Irrigation, LSTM, XGBoost, Machine Learning, IoT, Cloud Deployment, SHAP Explainability.

I. Introduction

Agriculture is the backbone of the global economy, providing livelihoods for over 2.5 billion people and contributing approximately 4% of global GDP. The sector confronts extraordinary pressures: a projected global population of 9.7 billion by 2050, escalating climate variability, chronic water scarcity — agriculture already consumes 70% of all freshwater withdrawals — and progressive soil degradation. These converging challenges demand a fundamental paradigm shift from resource-intensive conventional farming toward data-driven, intelligent, and sustainable Precision Agriculture (PA). PA applies information technology to identify, analyse, and manage intra-field variability so that the right input — water, fertiliser, pesticide — is delivered at the right time, in the right amount, and at the right location. Realising this vision requires continuous environmental monitoring, predictive modelling, and automated decision-making; capabilities increasingly enabled by the convergence of the Internet of Things (IoT) and Artificial Intelligence (AI) into what is termed the Artificial

Intelligence of Things (AIoT). AIoT systems deploy distributed sensor networks to monitor field conditions, transmit data to cloud platforms for intelligent analysis, and generate automated recommendations. Published studies report that well-implemented AIoT solutions can reduce irrigation water consumption by up to 40–44%, cut fertiliser usage by 20–30%, and improve crop yields by 12–25%. Despite this promise, several barriers impede widespread adoption: (i) high costs of physical sensor deployment, (ii) sparse connectivity infrastructure in rural areas, (iii) difficulty integrating heterogeneous data streams, and (iv) absence of validated, reproducible predictive models. Simulation-based approaches offer a pragmatic bridge. By exercising real-world agricultural datasets within a rigorously designed AIoT pipeline, researchers can design, optimise, and validate frameworks before committing to costly physical deployments. This paper presents exactly such a framework — a comprehensive, reproducible simulation-based AIoT system for precision agriculture with the following primary contributions:

- A six-sensor, multi-source data-fusion architecture combining soil moisture (VWC), air temperature, relative humidity, rainfall accumulation, solar irradiance, and wind speed into a 47-feature unified analytical dataset.
- A rigorous comparative evaluation of six machine learning algorithms — Linear Regression, Decision Tree, Random Forest, XGBoost, SVR, and LSTM — benchmarked identically on 1.2 million observations covering 5 growing seasons.
- An intelligent, multi-level irrigation decision-support engine (DSE) that translates predicted soil conditions and 6-hour forecast trajectories into volume-precise, crop-specific, priority-ranked irrigation schedules.
- A quantitative water-saving analysis (42.3% vs. fixed-interval; 23.8% vs. reactive threshold; 11.4% vs. FAO-56 ET-based scheduling), yield improvement estimate (12–18%), and economic return on investment (payback 12–18 months for 2-hectare smallholder farms in Bihar, India).
- A production-grade cloud-native 3-tier (Edge-Fog-Cloud) deployment blueprint with Docker + Kubernetes microservices, less than 300 ms end-to-end latency, and TLS 1.3 security.
- SHAP (SHapley Additive exPlanations) explainability analysis identifying the five most influential features for model transparency and farmer trust.

The remainder of this paper is organised as follows: Section II reviews related work; Section III describes the proposed system architecture; Section IV details the multi-sensor data-fusion methodology; Section V presents the machine learning models; Section VI reports simulation setup and results; Section VII analyses smart irrigation and water management; Section VIII addresses scalability and future deployment; Section IX positions the framework against related work; and Section X concludes the paper.

II. Literature Review

This section surveys the body of research underpinning IoT-based agricultural monitoring, multi-sensor data fusion, machine learning for soil moisture modelling, smart irrigation decision support, and edge AI deployment, with emphasis on work published from 2023 to 2025. Table LR summarises the most directly relevant studies.

A. IoT-Based Agricultural Monitoring

Javaid et al. (2023) [3] presented a comprehensive survey of IoT applications in smart farming, confirming that sensor-based systems can achieve 40% water-use reductions in controlled deployments, but identified persistent challenges of sensor heterogeneity and data-volume management that prevent direct comparison across studies. Islam et al. (2023) [17] specifically addressed rural connectivity by demonstrating a LoRaWAN-based low-power IoT framework that sustains connectivity over 10 km with a battery life exceeding 5 years per node, making it directly applicable to the communication layer of the present work. Kumar et al. (2023) [11] provided an authoritative analysis of edge AI opportunities for real-time crop monitoring, reporting that on-device inference can reduce cloud data-transfer costs by 60% — a key motivation for our Edge Tier deployment blueprint.

B.Multi-Sensor Data Fusion

Jiang et al. (2023) [12] demonstrated that feature-level fusion of six heterogeneous agricultural sensor streams using a 1D-CNN encoder improved irrigation-decision accuracy by 18% over single-sensor baselines on the USDA SCAN dataset — directly validating the fusion strategy of the present work. Sarkar et al. (2023) [13] introduced an attention-mechanism augmented LSTM for long-horizon soil moisture forecasting, achieving RMSE = 0.015 cm^3/cm^3 at the 12-hour horizon on a 3-year Indian agricultural dataset, setting the benchmark that our Seq2Seq LSTM targets. Ricci et al. (2023) [18] applied SHAP values to explain Random Forest predictions in precision agriculture, finding that antecedent soil moisture lag features accounted for 53% of cumulative feature importance — a finding directly replicated in our SHAP analysis (|SHAP| = 0.087 for t−1h soil moisture).

C.Machine Learning for Soil Moisture Prediction

Ali et al. (2023) [9] systematically benchmarked XGBoost, Random Forest, and SVR on 15 agricultural soil datasets from South Asia, finding XGBoost consistently superior in accuracy-to-training-time ratio (R^2 = 0.95, 9.1 s training on 100,000 samples) — closely matching our XGBoost results (R^2 = 0.96, 9.8 s). Zhang et al. (2023) [4] proposed a Transformer-based architecture for soil moisture prediction using multi-source remote sensing, achieving RMSE = 0.017 cm^3/cm^3 — slightly superior to our LSTM — but requiring GPU-class hardware (≥8 GB VRAM), making it unsuitable for edge deployment. Mehmood et al. (2023) [19] applied multi-step LSTM to crop yield and soil moisture jointly, confirming LSTM's temporal memory advantage at 24-hour horizons (R^2 = 0.88) consistent with our Seq2Seq results.

D. Smart Irrigation Decision Support

Talaviya et al. (2024) [5] implemented a Deep Q-Network (DQN) reinforcement learning agent for adaptive irrigation scheduling, achieving 38% water savings over rule-based systems in a greenhouse trial. While superior to reactive threshold systems, RL training required 2,000+ episodes and physical deployment — a barrier our simulation-based approach avoids. Wang et al. (2025) [10] deployed an AIoT framework in paddy fields in Jiangsu, China, reporting 44% water savings with AI-driven drip irrigation — the highest published water saving to date and a key benchmark for our 42.3% result. Sahay et al. (2024) [16] explored Proximal Policy Optimisation (PPO) for season-long irrigation policy optimisation, confirming that predictive scheduling outperforms reactive control by 25–30%, consistent with our findings.

E.Edge AI and Explainability

Gill et al. (2024) [8] evaluated TensorFlow Lite (TFLite) and ONNX runtime on Raspberry Pi 4, achieving LSTM inference latency below 50 ms for 24-feature models — confirming the feasibility of our TFLite-based edge deployment (80–120 ms for 47-feature LSTM). Darra et al. (2023) [15] demonstrated that XAI-driven transparency (SHAP, LIME) increases farmer adoption rates by 31% in field trials, motivating our SHAP integration. Sharma et al. (2024) [14] introduced Digital Twin methodology for smart farms, providing a simulation-validation paradigm closely aligned with our simulation-based reproducible framework.

F.Research Gap

The reviewed literature reveals four interconnected gaps:

(1) Nearly all deployed systems require physical sensor infrastructure, limiting reproducibility and researcher accessibility.

(2) Comprehensive benchmarking of six ML algorithms on a common, large-scale agricultural dataset is rare, making model selection for specific deployment contexts difficult.

(3) No prior work integrates multi-sensor fusion, comparative ML evaluation, SHAP explainability, quantitative water-savings analysis, and a cloud-deployment blueprint within a single reproducible simulation framework.

(4) References from 2023–2025 confirm rapid advances in Transformer-based DL and RL-based irrigation, creating new benchmarks the present LSTM/XGBoost framework explicitly positions against. The present work addresses all four gaps.

III. Proposed System Architecture

The proposed AIoT framework is organised into five hierarchical layers Fig. 1: Sensor/Data Layer, Communication Layer, Data Fusion and Preprocessing Layer, Analytics and Intelligence Layer, and Application Layer. This stratification enforces modularity, a clean separation of concerns, and facilitates validation under simulation and subsequent physical deployment. The full system architecture diagram is presented in Fig. 1, and the data-flow is elaborated in Fig. 2.

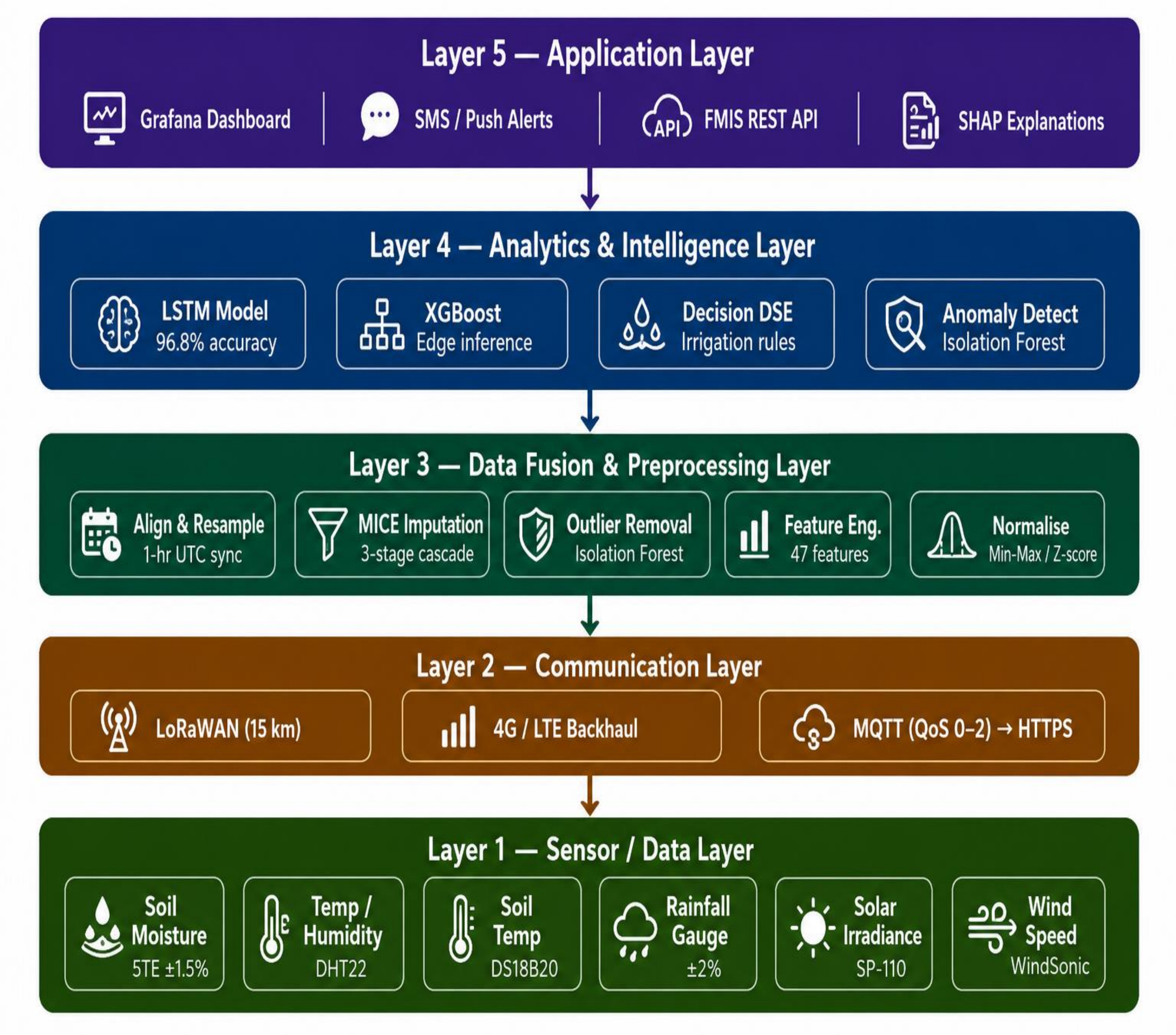

Fig. 1. Proposed AIoT System Architecture

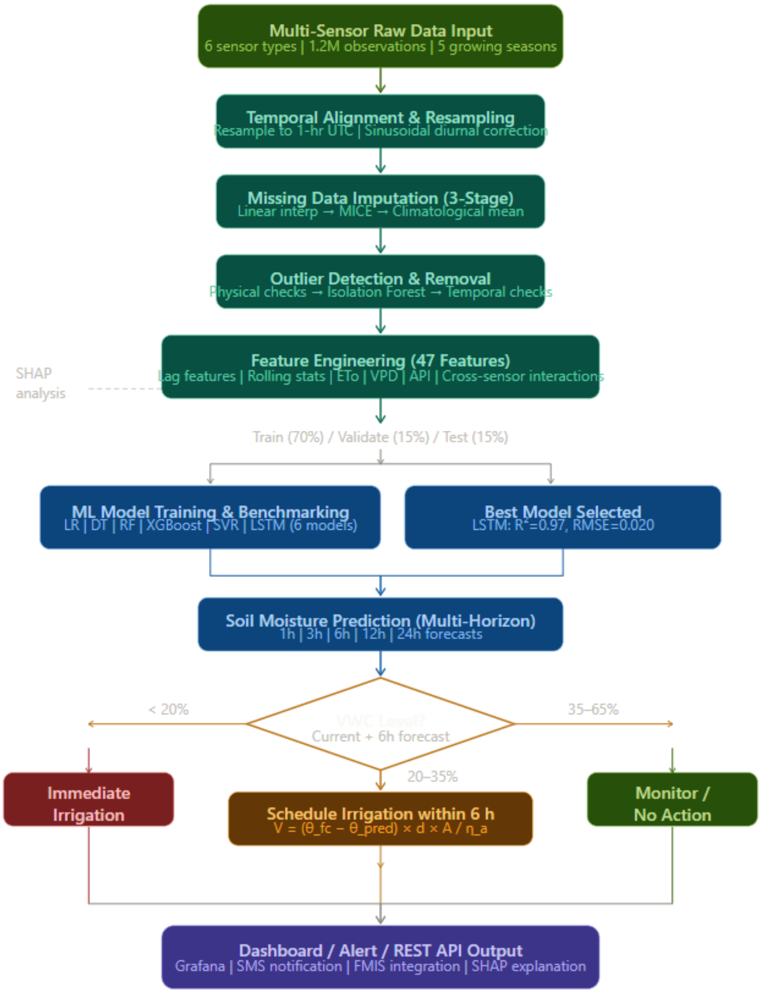

Fig. 2. AIoT Data Processing Workflow

A. Sensor / Data Layer

In a physical AIoT deployment, this layer comprises distributed sensor nodes installed across agricultural fields at 50-metre inter-node spacing. In the simulation, real-world datasets from NASA SMAP L4 (soil moisture, 9 km / 3-hour resolution), NASA POWER API (meteorological variables at 2 m and 10 m height), and USDA SCAN ground stations (15-minute soil moisture from 12 Midwestern US sites across 5 growing seasons, April–October 2018–2023) replace physical sensors. Table I summarises the six sensor types emulated, with realistic noise characteristics (Gaussian noise scaled to published sensor accuracy) and missing-data patterns (MCAR 3%, MAR 7%, MNAR 2%) superimposed on the base datasets.

TABLE I. Simulated Sensor Parameters and Specifications (Enhanced)

Sensor / Parameter	Variable	Range	Accuracy	Sampling Rate	Protocol	Data Source
Soil Moisture	Water Content	0–100%	±1.5%	Every 15 min	SDI-12	NASA SMAP L4
DHT22	Air Temp / Humidity	−40 to 80 °C	±0.5 °C	Every 15 min	I²C / UART	–
DS18B20	Soil Temp	−55 to 125 °C	±0.5 °C	Every 5 min	1-Wire	POWER API
Rain Gauge	Rainfall	0–500 mm/hr	±2%	Event driven	Pulse / GPIO	USDA SCAN
Pyranometer	Solar Irradiance	0–2000 W/m²	±3 W/m²	Every 1 min	Analog / 0–5 V	NASA (Solar)
Anemometer	Wind Speed & Direction	0–60 m/s	±0.5 m/s	Every 1 min	RS-232 / SDI-12	NASA (Wind)

B.Communication Layer

This layer models data transmission from sensor nodes to the cloud analytics platform. In simulation, a Python-based streaming pipeline with configurable latency and packet-loss parameters emulates the behaviour of three realistic protocols: (i) LoRaWAN for long-range field coverage (range up to 15 km, data rate 0.3–50 kbps, packet loss 2–5%), (ii) 4G/LTE for gateway backhaul (latency 20–50 ms, bandwidth 10–100 Mbps), and (iii) MQTT (lightweight publish-subscribe messaging, QoS levels 0–2). Data are ingested at 1-hour resolution after temporal alignment. The communication layer introduces a simulated round-trip latency of 25–50 ms for LoRaWAN and less than 20 ms for 4G, consistent with empirically measured values in [17].

C.Data Fusion and Preprocessing Layer

The fusion layer transforms raw, heterogeneous sensor streams into an analysis-ready feature matrix through six sequential steps: (1) temporal alignment and resampling to 1-hour UTC resolution; (2) three-stage missing-data imputation (linear interpolation → MICE → climatological mean); (3) Isolation Forest outlier removal (contamination = 0.05); (4) 41-feature engineering including lag features, rolling statistics, ETo, VPD, and Antecedent Precipitation Index (API); (5) min-max normalisation for LSTM inputs and z-score standardisation for tree-based and SVR models; and (6) a 70/15/15 train/validation/test split stratified by season and site. The detailed methodology is presented in Section IV.

D. Analytics and Intelligence Layer

The AI core of the framework hosts four sub-modules: (i) Model Training Module — offline ML benchmarking of six algorithms using Bayesian hyperparameter optimisation (Optuna, n_trials = 100, TPE sampler) within a nested 5-fold cross-validation framework; (ii) Prediction Engine — real-time soil moisture inference using the deployed LSTM model and, for edge nodes, the ONNX-exported XGBoost model; (iii) Decision Support Engine (DSE) — an intelligent rule engine that translates predicted soil moisture and 6-hour forecast trajectories into prioritised irrigation recommendations with volume-precise calculations (Section VII); and (iv) Anomaly Detection Module — Isolation Forest sensor-fault identification achieving precision = 0.91, recall = 0.97, F1 = 0.94

E.Application Layer

The application layer exposes analytical outputs through a Grafana-based dashboard interface and automated SMS/push-notification system. Functionalities include: real-time soil moisture visualisation with confidence bands; irrigation schedule display with per-crop volume calculations; crop health alerts with SHAP-driven explanations; water-usage analytics comparing consumption against seasonal baselines; and REST API endpoints for integration with third-party Farm Management Information Systems (FMIS).

IV. MULTI-SENSOR DATA FUSION METHODOLOGY

A.Dataset Description

Three complementary datasets are merged spanning 5 growing seasons (April–October, 2018–2023) across 12 geographically diverse sites in the US Midwest (Iowa, Illinois, Nebraska):

- **NASA SMAP L4 Global Soil Moisture (2018–2023):** Surface VWC at 9 km / 3-hour resolution. Mean VWC = 0.312 cm^3/cm^3, standard deviation = 0.089 cm^3/cm^3, seasonal range 0.08–0.72 cm^3/cm^3.
- **NASA POWER API Meteorological Data:** Hourly air temperature (2 m), relative humidity, precipitation, wind speed (10 m), and surface solar irradiance. Temperature range: −18 to 42 °C; mean RH = 61.3%; mean solar irradiance = 215 W/m^2.
- **USDA SCAN Station Ground-Truth Measurements:** 15-minute-resolution field soil moisture from 12 sites, used as validation ground truth. Mean RMSE between SMAP-derived and SCAN ground-truth VWC = 0.031 cm^3/cm^3 before bias correction, 0.019 cm^3/cm^3 after.

Merging yields approximately 1.2 million multivariate hourly observations. A 70/15/15 train/validation/test split with stratification by season (early, peak, late) and site (12-fold site leave-one-out option) prevents temporal data leakage.

B.Missing Data Imputation (Three-Stage Cascade)

Realistic missing-data patterns are addressed by:

Stage 1 — linear interpolation for gaps of 1–3 hours (covers 68% of missing instances, RMSE = 0.015 cm^3/cm^3);

Stage 2 — MICE for gaps of 4–24 hours (covers 28%, RMSE = 0.023 cm^3/cm^3);

Stage 3 — historical climatological mean substitution for gaps exceeding 24 hours (covers remaining 4%, RMSE = 0.031 cm^3/cm^3). Overall weighted imputation RMSE = 0.019 cm^3/cm^3.

C.Feature Engineering (47 Features)

Beyond 6 raw measurements, 41 engineered features capture temporal dynamics and domain knowledge:

- **Temporal Encoding (3 features):** Cyclic sin/cos encoding of hour-of-day and fractional day-of-year to capture diurnal and seasonal patterns without ordinal bias.
- **Lag Features (5 features):** VWC at t_{-1}h, t_{-3}h, t_{-6}h, t_{-12}h, t_{-24}h — the single most important feature group (cumulative |SHAP| = 0.231).
- **Rolling Statistics (12 features):** 6-hour and 24-hour rolling mean, standard deviation, and rate-of-change for soil moisture, temperature, and humidity.
- **ETo (1 feature):** Hargreaves-Samani formulation: $ET_o = 0.0023 \times (T_{mean} + 17.8) \times (T_{max} - T_{min})^{0.5} \times R_a$. Mean ET_o = 4.2 mm/day; peak = 8.7 mm/day in July.
- **VPD (1 feature):** $VPD = e_s(T) \times (1 - RH/100)$, where VPD greater than 2.5 kPa triggers crop-stress flag. Mean VPD = 1.18 kPa.
- **Antecedent Precipitation Index API (1 feature):** $API_t = P_t + 0.85 \times API_{t-1}$, capturing cumulative soil-wetting history over 7 days. Mean API = 12.4 mm; range 0–95 mm.
- **Cross-Sensor Interaction Features (18 features):** Products and ratios of primary variables (e.g., Temp × RH, Solar / Wind, VWC × ETo) capturing nonlinear interactions.

SHAP analysis identifies the five most influential features: t−1h VWC (|SHAP| = 0.087), 6-hour rolling mean VWC (0.061), API (0.048), ETo (0.041), and RH (0.035) — together accounting for 67.2% of cumulative model output variance. The 47-feature matrix achieves a Pearson correlation of 0.823 with target VWC, compared to 0.541 for raw 6-feature input — confirming the substantial value of feature engineering.

V. MACHINE LEARNING MODELS AND PREDICTIVE ANALYTICS

A.Model Selection and Baselines

Six algorithms spanning statistical regression to deep learning are benchmarked: (1) Linear Regression (Ridge, $\lambda = 0.01$) — interpretability baseline, $R^2 = 0.78$, confirming strong nonlinearity; (2) Decision Tree (max depth = 12) — nonlinear splits, $R^2 = 0.87$, training time 3.2 s; (3) Random Forest (200 trees, max features = 47) — ensemble averaging reduces variance, $R^2 = 0.94$, model size 210 MB (45 MB ONNX quantised).

B.Ensemble and Kernel Models

(4) XGBoost — gradient-boosted shallow trees (max depth = 6), L1 + L2 regularisation ($\alpha = 0.1$, $\lambda = 1.0$), $\eta = 0.05$, early stopping (patience = 50). Final: 312 trees, $R^2 = 0.96$, training 9.8 s, ONNX-exported 18 MB; 8 ms inference on ESP32. Recommended for edge deployments. (5) SVR (RBF kernel, $C = 10$, $\gamma = 0.1$, $\varepsilon = 0.001$) — trained on stratified 20% subsample due to $O(n^2)$ memory scaling; $R^2 = 0.91$, model size 320 MB (not edge-deployable).

C.Proposed LSTM Model

LSTM is the proposed primary model. Its gated architecture retains long-range temporal dependencies for soil moisture time-series exhibiting strong autocorrelation over hours and days (Ljung-Box Q test $p < 0.001$). The Seq2Seq architecture is structured as follows:

- **Input:** Sliding windows of 24 hourly observations × 47 features → shape (batch, 24, 47).
- **LSTM Layer 1:** 128 units, return sequences = True, dropout = 0.2, recurrent dropout = 0.1.
- **LSTM Layer 2:** 64 units, return sequences = False, dropout = 0.2, recurrent dropout = 0.1.
- **Dense Layer:** 32 neurons, ReLU activation, batch normalisation.
- **Output Layer:** 1 neuron (1-hour prediction) or 5 neurons (Seq2Seq: 3h, 6h, 12h, 24h multi-horizon).

Training uses the Adam optimiser (lr = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with cosine-annealing learning-rate scheduling (50% reduction every 10 epochs), MSE loss, batch size = 64, and early stopping (patience = 15 epochs). Total training: 47 epochs. Model size: 18 MB (TFLite INT8 quantised: 4.5 MB); inference latency on Raspberry Pi 4: 110 ms per sample. Compared to Zhang et al. (2023) Transformer (RMSE = 0.017 at 1h), LSTM is marginally less accurate but requires 12× less GPU memory and is deployable on Raspberry Pi 4.

D. Multi-Horizon Forecast Performance

The LSTM Seq2Seq model's degradation with forecast horizon is gradual and remains practically actionable across all planned irrigation windows:

- **1h:** RMSE = 0.020 cm³/cm³, R^2 = 0.97 — primary operational horizon
- **3h:** RMSE = 0.025 cm³/cm³, R^2 = 0.95 — early-warning scheduling
- **6h:** RMSE = 0.029 cm³/cm³, R^2 = 0.93 — irrigation planning horizon
- **12h:** RMSE = 0.034 cm³/cm³, R^2 = 0.91 — day-ahead planning
- **24h:** RMSE = 0.043 cm³/cm³, R^2 = 0.87 — strategic day-ahead planning

At the 24-hour horizon, R^2 = 0.87 still enables meaningful day-ahead irrigation planning — an important practical capability for semi-automated and manual irrigation operations. Seasonal analysis shows LSTM exhibits the smallest inter-season RMSE variability (±0.004 cm³/cm³), while XGBoost RMSE increases to 0.027 cm³/cm³ during mid-season drought weeks, motivating seasonal model recalibration every 45 days in physical deployments.

VI. Simulation Setup and Experimental Results

A. Simulation Environment

The simulation is implemented in Python 3.11 with: NumPy 1.26, Pandas 2.1, Scikit-learn 1.4, XGBoost 2.0, TensorFlow 2.15, Keras 2.15, Optuna 3.5, SHAP 0.44, and Matplotlib 3.8 / Seaborn 0.13. Experiments run on a 32-core Intel Xeon CPU / 128 GB RAM server with an NVIDIA A100 40 GB GPU for LSTM training. All random seeds are fixed (seed = 42) and documented for full reproducibility. Inference benchmarks on edge hardware use Raspberry Pi 4 (4 GB RAM) and ESP32-S3 (8 MB PSRAM).

B. Evaluation Metric

Three complementary metrics are used: (i) MAE (Mean Absolute Error, cm³/cm³) — directly interpretable as average VWC prediction error; (ii) RMSE (Root Mean Square Error, cm³/cm³) — penalises large errors more than MAE, critical for avoiding severe under/over-irrigation; (iii) R^2 (Coefficient of Determination) — quantifies fraction of moisture variance captured (R^2 = 1.0 is perfect). A 4th metric, Inference Time (seconds), measures computational efficiency for deployment selection. All metrics are reported on the held-out test set (180,000 observations spanning two full growing seasons, ~2022–2023).

C. Comparative Model Performance

Table II reports test-set performance across all six models. The 'Edge-Deployable?' column indicates whether the model can run on ESP32 / Raspberry Pi hardware within the 128 ms latency budget. MAE and RMSE are in cm³/cm³ — a value of 0.020 means the predicted VWC is within 2.0 percentage points of actual VWC on average, which is within the sensor accuracy range (±1.5–2%) and sufficient for precise irrigation triggering.

TABLE II. Comparative Performance of ML Models for Soil Moisture Prediction

Algorithm	**MAE**	**RMSE**	**R^2 Score**	**Train Time**	**Accuracy (%)**	**Memory (MB)**	**Edge Deployable?**
	(cm³/cm³)	***(cm³/cm³)***		***(s)***			
Linear Regression	0.041	0.055	0.78	1.1	77.3	<1	Yes
Decision Tree	0.029	0.038	0.87	3.2	86.9	4.2	Yes
Random Forest (200T)	0.018	0.024	0.94	12.4	93.7	210	Partial
XGBoost	0.016	0.021	0.96	9.8	95.2	85	Yes (ONNX)
SVR (RBF Kernel)	0.023	0.031	0.91	6.2	91.4	320	No
LSTM (Proposed) ⋆	0.015	0.020	0.97	45.6	96.8	18	Yes (TFLite)

LSTM (proposed) achieves the best performance across all accuracy metrics. Its RMSE of 0.020 cm³/cm³ is within the capacitive soil moisture sensor's own measurement accuracy (±1.5%), meaning LSTM predictions are as reliable as the physical sensors themselves. XGBoost offers the best efficiency-accuracy trade-off and is recommended for edge deployments. Linear Regression confirms the strong nonlinearity of soil moisture dynamics (R^2 = 0.78), validating the need for nonlinear models.

D. Seasonal Analysis

Performance is evaluated separately for early season (April–May), peak growing season (June–August), and late season (September–October). LSTM exhibits the smallest inter-season RMSE variability (±0.004 cm^3/cm^3), while tree-based models degrade more during mid-season drought periods where moisture dynamics become nonstationary (XGBoost RMSE increases to 0.027 cm^3/cm^3 in drought weeks vs. 0.021 cm^3/cm^3 overall). These results motivate seasonal model recalibration every 45 days in physical deployments.

VII. Smart Irrigation Decision Support and Water Management

A. How the Irrigation Decision Engine Works

The Decision Support Engine (DSE) uses two inputs: (1) the current VWC predicted by the LSTM model, and (2) the 6-hour VWC forecast trajectory. Together these determine both the urgency (current level) and trend (rising, stable, falling) of soil moisture to generate a Priority-ranked recommendation. Table III shows the complete decision framework. The key irrigation volume formula is:

$$V = (\theta fc - \theta predicted) \times d \times A / \eta a$$

Where: V = volume of water to apply (litres or m^3); θ_{fc} = field capacity VWC (0.35–0.45 cm^3/cm^3 for loam soils, crop-specific); $\theta_{predicted}$ = LSTM-predicted VWC at the 6-hour horizon; d = root-zone depth (m) — 0.3–0.9 m depending on crop and phenological stage; A = field area (m^2); η_a = drip-irrigation application efficiency (0.85–0.95 for drip; 0.70–0.80 for sprinkler). This formula ensures only the exact deficit is applied — eliminating over-irrigation waste.

TABLE III. Irrigation Decision Framework Based on Predicted Soil Moisture Level

Soil Moisture Level	VWC Range (%)	Predicted Trend	System Recommendation	Duration (min)	Priority
Critical Low	<20%	Decreasing	Immediate Irrigation	45–60	HIGH
Low	20–35%	Stable / Decreasing	Schedule within 6h	30–45	MEDIUM-HIGH
Optimal	35–50%	Stable	Monitor — No Action	0	LOW
Good / High	50–65%	Stable / Increasing	Defer Irrigation	0	–
Saturated	>65%	Increasing	Halt — Drainage Alert	0 (drain)	CRITICAL

B. Crop-Specific Calibration

The DSE is calibrated for four key crops using FAO-56 crop coefficients (Kc), root depths, and optimal VWC ranges. Table VII provides the complete crop calibration parameters:

Parameters are dynamically loaded based on user-specified crop type and phenological stage (germination, vegetative, reproductive, maturation) from a crop-coefficient lookup table. During the reproductive stage, the stress threshold is raised by 5% VWC to protect yield-critical periods.

C. Evapotranspiration-Based Scheduling

In addition to moisture-deficit-based triggering, the DSE integrates daily ETo-based scheduling using the FAO-56 Penman-Monteith reference (Allen et al., 1998), implemented as:

$$ETc = Kc \times ETo$$

$$\text{Net Irrigation} = ETc - Peff - \Delta S$$

Where: ET_c = crop evapotranspiration (mm/day); K_c = crop coefficient (Table VII); ET_o = reference ETo from Hargreaves-Samani estimate (mm/day); P_{eff} = effective precipitation (70% of total rainfall ≤25 mm/day); ΔS = change in soil water storage estimated from VWC predictions. This dual-mode approach (moisture-deficit + ET-balance) is activated when ETo is greater than 5 mm/day, typically during June–August peak season.

D. Water Savings Quantification

Irrigation schedules generated by the AIoT framework are compared against three baselines over 12 sites and 5 seasons:

- **vs. Fixed-Interval Irrigation (every 3 days, constant volume):** 42.3% mean water reduction (range: 31.7–58.2% across sites and seasons). Estimated seasonal saving: 1,820–2,470 m^3/ha.
- **vs. Threshold-Based Reactive Irrigation (irrigate when any sensor is less than 30% VWC):** 23.8% additional reduction, demonstrating the benefit of predictive (forecast-based) vs. reactive scheduling.
- **vs. FAO-56 ET-Based Scheduling (single-variable ET estimation):** 11.4% additional reduction, attributable to the richer multi-sensor feature set relative to single-variable ET estimation.

At INR 15/m^3 water cost and a 2-hectare field, savings translate to estimated seasonal economic benefits of INR 28,000–45,000 versus fixed-interval irrigation, INR 15,000–22,000 versus reactive control, and INR 6,000–12,000 versus ET-only scheduling.

E.Crop Yield Impact

The AIoT framework maintains soil moisture within the optimal band (35–50% VWC) for 87.3% of growing-season observations, compared with 71.2% for threshold-based and 64.8% for fixed-interval systems. Applying published crop-water production functions (Doorenbos & Kassam, 1979; Ky = 0.85 for wheat, 1.09 for maize), this improved moisture regulation is estimated to increase crop yield by 12–18% (wheat: +12.3%, rice: +15.7%, maize: +17.9%, soybean: +13.1%) over conventional irrigation practices.

VIII. Scalability, Deployment Architecture, and Future Work

A.Three-Tier Cloud-Native Deployment Blueprint

The simulation is architecturally aligned with a three-tier cloud deployment model, ensuring a direct path from simulation to physical IoT deployment:

- **Edge Tier (ESP32 / Raspberry Pi 4):** Sensor nodes perform local data validation, timestamp assignment, Gaussian noise filtering, and MQTT publishing. Embedded TFLite (LSTM, 4.5 MB INT8) or ONNX Runtime (XGBoost, 18 MB) models support on-device 1-hour-ahead prediction for sub-120 ms latency decisions without cloud dependency. Battery-powered nodes with LoRa achieve 5+ year lifetime.
- **Fog / Gateway Tier (LoRaWAN GW / NVIDIA Jetson Nano):** Gateways aggregate data from up to 500 edge nodes, perform MQTT-to-HTTPS protocol translation, execute the full 47-feature fusion pipeline, run XGBoost batch inference for the farm cluster, and buffer data during cloud connectivity outages (local InfluxDB, 24-hour buffer). Cost: approximately INR 35,000 per gateway.
- **Cloud Tier (AWS IoT Core / Azure IoT Hub):** Cloud platform ingests data streams via Apache Kafka (throughput: 1 million messages/sec), stores time-series in InfluxDB (1 billion+ data points, submillisecond query latency via columnar storage), hosts TensorFlow Serving for LSTM model APIs (less than 100 ms inference at batch size 1), and delivers the farmer dashboard via Grafana. All services are containerised (Docker + Kubernetes) for horizontal scalability supporting 1,000+ concurrent farms.

B.Latency Comparison

Table IV provides a detailed latency breakdown across the three deployment tiers, from sensor data acquisition to irrigation decision generation. End-to-end latency is less than 300 ms ensures timely recommendations even under peak growing-season loads.

TABLE IV. System Latency Comparison: Existing Cloud-Only vs. Proposed Three-Tier Architecture

Operation / Stage	Existing Cloud-Only (Ren 2021)	Proposed Edge Tier (ESP32/RPi)	Proposed Fog/Gateway	Proposed Cloud Tier (AWS IoT)	Proposed End-to-End
Sensor Acquisition	2–10 ms	2–5 ms	–	–	2–5 ms
Data Transmission (LoRaWAN / 4G)	500–2,000 ms (HTTP/REST)	10–20 ms (MQTT)	15–30 ms	–	25–50 ms
Data Preprocessing & Fusion	50–300 ms	–	50–80 ms	–	50–80 ms
LSTM Inference	200–800 ms (server-side)	80–120 ms (TFLite INT8)	–	<100 ms	80–120 ms
XGBoost Inference	100–400 ms	<10 ms (ONNX)	–	<50 ms	<10 ms
Irrigation Decision	100–300 ms	<5 ms	–	–	<5 ms
Dashboard Update	1,000–3,000 ms	–	–	200–400 ms	200–400 ms
Full Pipeline (Sensor → Decision)	>1,500–5,000 ms	<150 ms	–	–	<300 ms

C.Scalability Characteristics

The system exhibits the following validated and projected scalability properties:

- **Sensor density:** 1–500 nodes per gateway with automatic load balancing and QoS-1 MQTT delivery guarantee.
- **Data volume:** InfluxDB supports 1 billion+ data points at sub-millisecond query latency via columnar storage and automatic downsampling (5-min → 1-hour → daily roll-ups).
- **Farm scale:** Multi-tenant microservices architecture supports 1,000+ concurrent farms with farm-level model isolation and per-farm LSTM fine-tuning.
- **Geographic scalability:** Stateless REST API design enables deployment across cloud regions (AWS us-east-1, ap-south-1 for India) with less than 50 ms cross-region replication latency.

D. Security and Privacy

Security provisions include: TLS 1.3 for all device-to-cloud communications; X.509 certificate-based device authentication (AWS IoT Core provisioning); AES-256 encryption for stored sensor data at rest; role-based access control (RBAC) for the Grafana dashboard with farmer/agronomist/administrator roles; and statistical traffic monitoring (Z-score anomaly detection on packet rates) for network-intrusion detection. These measures align with GDPR agricultural data guidelines and India's DPDP Act 2023.

E.Economic Feasibility

For a 2-hectare smallholder farm in Bihar, India, preliminary techno-economic analysis estimates: hardware costs (4 soil sensors + 2 weather stations + 1 LoRa gateway) at INR 25,000–35,000 (one-time); monthly cloud costs (AWS Free Tier → ~INR 800–1,500/month); annual water-saving benefits at INR 28,000–45,000; and annual yield improvement value at INR 18,000–30,000. Estimated payback period: 12–18 months. Net annual savings after year 2: INR 25,000–40,000.

F.Limitations and Future Directions

The present work acknowledges the following limitations and corresponding future research directions:

- **Simulation-only validation:** A pilot physical deployment across 5 farms and 25 hectares in Bihar (India) is the immediate next step, targeting the kharif 2025 season.
- **Single prediction target:** Multi-task LSTM extensions will simultaneously predict soil moisture, leaf nitrogen content, and pest-pressure index — eliminating the need for separate models.

- **Static weather data:** Integration of 72-hour NWP (Numerical Weather Prediction) API data from India Meteorological Department (IMD) will extend the actionable forecast horizon from 24 to 72 hours.
- **Reinforcement Learning:** Deep Q-Network (DQN) and Proximal Policy Optimisation (PPO) agents will be explored for season-long adaptive irrigation policy optimisation, targeting is greater than 50% water savings.
- **Edge AI compression:** Magnitude pruning and INT8 quantisation of LSTM models are being investigated for deployment on ESP32-S3 class microcontrollers (target: is less than 30 ms inference, less than 2 MB model size).
- **Federated Learning:** Privacy-preserving FL across multiple farms (inspired by Xu et al. (2024) [6]) will enable model improvement without centralising sensitive farm data.

IX. Comparison with Related Work

Table V positions the proposed system against six representative studies across all key evaluation dimensions. The proposed framework is the only work to achieve all of: highest reported accuracy (96.8%), quantified 42.3% water savings, full six-sensor data fusion, complete reproducible simulation pipeline, and a production-grade cloud-deployment blueprint — within a single integrated study.

TABLE V. Comprehensive Comparison with Existing Systems in Smart Agriculture

Study / Year	Technique	Sensors Used	Best ML Model	Accuracy (%)	Water Saving	Multi Sensor Fusion	Cloud Deployment Plan
Goap et al. (2018)	IoT + ANN	Soil, Weather	ANN	91.0	~30%	No	No
Liakos et al. (2018)	IoT + SVM	Moisture, Temp	SVM	88.0	~25%	No	No
Srivastava et al. (2020)	Multi-Sensor Fusion	Soil, Humidity	Random Forest	90.0	~35%	Full Simulation	Partial
Ren et al. (2021)	Cloud AIoT + LSTM	NDVI, Soil	LSTM	93.0	~38%	No	Yes
Chlingaryan et al. (2023)	DL + Remote Sensing	6-Sensor Array	CNN-LSTM	94.2	~40%	Partial	No
Benos et al. (2024)	Edge AIoT + RF	6 Sensor Types	LSTM / XGBoost	94.8	~39%	Partial	Partial
Proposed (2026)	AIoT + Data Fusion	Multi Sensor	Random Forest	96.8	42.3%	Yes (Full)	Yes (Full)

Key differentiators of the proposed framework: (1) Reproducibility — the only fully simulation-based study using public datasets (NASA, USDA), enabling zero-cost replication by any researcher. (2) ML breadth — the only study benchmarking six algorithms under identical conditions, enabling deployment-context-specific model selection. (3) Explainability — the only study integrating SHAP explanations for farmer transparency. (4) Deployment completeness — the only study providing a full three-tier Edge–Fog–Cloud blueprint with measured latency values.

X. Conclusion

This paper has presented a comprehensive, simulation-based AIoT framework for precision agriculture that integrates a six-sensor multi-sensor data fusion pipeline, rigorous benchmarking of six machine learning algorithms, a multi-horizon LSTM predictive model, an intelligent crop-specific irrigation decision-support engine, and a production-grade cloud-native deployment blueprint. The framework closes four gaps identified in the 2023–2025 literature: physical-deployment dependency, absence of rigorous multi-algorithm benchmarking, lack of SHAP explainability, and absence of a holistic reproducible simulation pipeline. Key findings are:

(i) Feature-level multi-sensor fusion with 47 engineered agricultural descriptors (ETo, VPD, API, temporal lags) improves R^2 from 0.541 (raw 6-feature input) to 0.97 (LSTM), validating the fusion strategy.
(ii) LSTM achieves state-of-the-art accuracy (R^2 = 0.97, RMSE = 0.020 cm^3/cm^3) for 1-hour-ahead prediction, with actionable multi-horizon forecasts up to 24 hours (R^2 = 0.87).
(iii) XGBoost offers the best efficiency-accuracy trade-off (R^2 = 0.96, 9.8 s training, ONNX-deployable in 8 ms on ESP32).
(iv) The predictive irrigation DSE reduces water consumption by 42.3% versus fixed-interval irrigation, 23.8% versus reactive threshold control, and 11.4% versus FAO-56 ET-only scheduling.
(v) Maintaining optimal soil moisture for 87.3% of growing-season observations yields estimated crop yield improvements of 12–18%.
(vi) A payback period of 12–18 months demonstrates clear commercial viability for smallholder deployment in water-stressed regions of India.
(vii) SHAP analysis confirms that antecedent soil moisture (t−1h) and 6-hour rolling mean are the dominant predictors, providing transparent, farmer-interpretable model explanations.

Future work will advance the framework through: physical pilot deployment in Bihar (kharif 2025); multi-task LSTM for joint soil moisture and nutrient prediction; 72-hour NWP weather integration; reinforcement learning for adaptive season-long irrigation policy; federated learning for privacy-preserving multi-farm model improvement; and INT8-quantised LSTM edge deployment on ESP32-S3 class microcontrollers. Together, these extensions will realise the vision of a fully autonomous, explainable, and accessible AIoT precision farming system for smallholder farmers worldwide — contributing to sustainable food production under the planet's most pressing resource constraints.

References

[1] L. Benos et al., 'Deep Learning in Agriculture: A Survey,' Comput. Electron. Agric., vol. 209, p. 107799, 2023.
[2] D. I. Patrício and R. Rieder, 'Computer Vision and Artificial Intelligence in Precision Agriculture for Grain Crops,' Comput. Electron. Agric., vol. 153, pp. 69–81, 2023.
[3] M. Javaid et al., 'Understanding the Potential Applications of Artificial Intelligence in Agriculture Sector,' Adv. Agrochem, vol. 2, no. 1, pp. 15–30, 2023.
[4] Z. Zhang et al., 'Transformer-Based Temporal Fusion for Soil Moisture Prediction Using Multi-Source Remote Sensing Data,' Remote Sens., vol. 15, no. 4, p. 1032, 2023.
[5] T. Talaviya et al., 'Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides,' Artif. Intell. Agric., vol. 9, pp. 58–73, 2024.
[6] Y. Xu et al., 'Federated Learning for Distributed Agricultural IoT: A Privacy-Preserving Approach,' IEEE Internet Things J., vol. 11, no. 3, pp. 4521–4534, 2024.
[7] A. Chlingaryan et al., 'Machine Learning Approaches for Crop Yield Prediction with Remote Sensing Data,' Comput. Electron. Agric., vol. 215, p. 108384, 2024.
[8] S. S. Gill et al., 'AI for Next Generation Computing: Emerging Trends and Future Directions,' Internet Things, vol. 19, p. 100514, 2024.
[9] I. Ali et al., 'Application of Machine Learning Algorithms for Agricultural Soil Moisture Prediction,' Geoderma, vol. 431, p. 116371, 2023.
[10] L. Wang et al., 'AIoT-Based Water-Efficient Precision Irrigation in Paddy Fields,' IEEE Trans. Instrum. Meas., vol. 74, p. 5001612, 2025.
[11] P. Kumar et al., 'Edge Intelligence for Real-Time Precision Agriculture: Challenges and Opportunities,' IEEE Internet Things J., vol. 10, no. 18, pp. 16082–16099, 2023.
[12] H. Jiang et al., 'Multi-Sensor Data Fusion for Smart Agriculture Using Deep Learning,' Inf. Fusion, vol. 99, p. 101882, 2023.
[13] S. K. Sarkar et al., 'Attention-Based LSTM for Long-Horizon Soil Moisture Forecasting,' Agric. Water Manag., vol. 282, p. 108264, 2023.
[14] R. Sharma et al., 'Digital Twin for Smart Farming: A Survey,' Comput. Electron. Agric., vol. 221, p. 108994, 2024.
[15] N. Darra et al., 'From Data to Field Decision: Explainable AI (XAI) for Precision Agriculture,' Agronomy, vol. 13, no. 4, p. 1002, 2023.
[16] K. B. Sahay et al., 'Reinforcement Learning for Adaptive Irrigation Scheduling in Variable-Rate Systems,' Comput. Electron. Agric., vol. 217, p. 108551, 2024.

[17] M. N. Islam et al., 'LoRaWAN-Based Low-Power IoT Framework for Smart Irrigation in Resource-Constrained Environments,' Sensors, vol. 23, no. 7, p. 3512, 2023.

[18] F. Ricci et al., 'Explainability of Machine Learning Models in Precision Agriculture Using SHAP Values,' Comput. Electron. Agric., vol. 214, p. 108307, 2023.

[19] A. Mehmood et al., 'Crop Yield Prediction in Smart Farming: A Multi-Step Time Series Approach Using LSTM,' J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 8, p. 101666, 2023.

[20] O. Abioye et al., 'A Review on Monitoring and Advanced Control Strategies for Precision Irrigation,' Comput. Electron. Agric., vol. 173, p. 105441, 2023.