

Machine Learning for Crop Yield Prediction Using Sensor Data

Mr. P. Balamurugan

Assistant Professor,
Department of Computer Science and Engineering,
Kings College of Engineering,
Punalkulam, TN, India,
p.balucse83@gmail.com

Hariharan. N

Department of Computer Science and Engineering,
Kings College of Engineering,
Punalkulam, TN, India,
hariharanhari102004@gmail.com

ABSTRACT

Crop yield prediction is central to agricultural planning, food security, and farm-level decision-making, yet conventional approaches still rely heavily on manual field estimation and the subjective experience of farmers or agronomists. Such methods struggle to incorporate real-time environmental variability, are slow to update, and cannot easily reconcile the many interacting factors that govern plant growth, including soil moisture, temperature, humidity, rainfall, and nutrient availability. The proliferation of low-cost Internet of Things (IoT) sensors has made it possible to continuously capture fine-grained soil and atmospheric parameters directly from the field, creating an opportunity to replace episodic manual assessment with data-driven, near-real-time prediction. This paper proposes a machine-learning-based framework for crop yield prediction that ingests multi-parameter sensor data — including soil moisture, soil and air temperature, relative humidity, rainfall, soil pH, electrical conductivity, and NPK nutrient levels — and transforms it into yield estimates through a structured pipeline of data acquisition, cleaning, feature engineering, and model training. Six representative learning algorithms are examined and compared conceptually: Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Gradient Boosting/XGBoost, and Artificial Neural Networks, each evaluated for its suitability to the nonlinear, noisy, and heterogeneous nature of agricultural sensor data. Standard regression metrics — Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and the coefficient of determination (R^2) — are described as the basis for model comparison. As no field deployment or labelled dataset was collected for this study, the work is presented as a proposed framework and illustrative experimental design rather than a report of empirical results, and all performance figures shown are clearly marked as illustrative placeholders for the evaluation structure that a full implementation would populate. The discussion highlights the practical significance of the framework for farmers and agricultural planners, including support for irrigation scheduling, fertilizer optimization, and early identification of yield decline, while also outlining the limitations of sensor-based prediction systems, such as calibration drift, missing data, and regional variability, and the future directions needed to move from prototype to field-ready deployment.

Keywords: Machine Learning, Crop Yield Prediction, Agricultural IoT, Sensor Data, Precision Agriculture, Smart Farming, Regression, Crop Monitoring

I. INTRODUCTION

A. Background of Agriculture and Crop Yield Prediction

Agriculture remains the foundation of food security and rural livelihoods across the world, and the ability to anticipate how much a given crop will yield before harvest is essential for planning at every scale, from an individual farmer deciding how much fertilizer to apply, to a national agency planning grain reserves. Crop yield is the outcome of a large number of interacting factors rather than any single variable. Climate variability determines the temperature and rainfall regime a crop experiences over its growth cycle; soil conditions such as texture, organic content, and nutrient status determine how effectively a plant can take up water and minerals; irrigation timing and volume influence water stress; and pest or disease pressure can sharply reduce the realized yield relative to its potential. Because these factors vary continuously in time and space, even within a single field, accurate yield prediction requires continuous, multi-parameter observation rather than a single end-of-season estimate.

B. Limitations of Conventional Crop Yield Prediction

Traditional yield estimation is still performed largely through manual field visits, visual inspection, and the accumulated experience of farmers or agricultural extension officers. While this experiential knowledge is valuable, it does not scale well and is difficult to standardize across regions or seasons. Manual estimation is time-consuming, is typically performed at coarse time intervals, and provides limited real-time information about how a crop is responding to short-term stress such as a dry spell or a heat wave. Because manual observation cannot easily reconcile many variables simultaneously, decisions are frequently made on the basis of a small number of visible cues, leaving important underlying signals such as soil moisture depletion or early nutrient deficiency undetected until visible symptoms appear. This delay in decision-making, combined with inconsistent measurement practices between observers, contributes to yield forecasts that are frequently inaccurate, particularly under increasingly unpredictable climate conditions.

C. Role of IoT Sensors in Agriculture

The falling cost of Internet of Things (IoT) sensing hardware has made continuous, automated environmental monitoring practical even for small and medium-sized farms. Soil moisture probes, soil and air temperature sensors, relative-humidity sensors, tipping-bucket rain gauges, light intensity (LDR) sensors, soil pH probes, electrical-conductivity (EC) sensors, and NPK nutrient sensors can now be deployed as a distributed network across a field, transmitting readings at regular intervals to a local gateway and onward to cloud storage. Unlike periodic manual sampling, this arrangement produces a dense, time-continuous record of the conditions a crop actually experiences, capturing short-lived events such as a sudden drop in soil moisture or an unseasonal temperature spike that would otherwise be missed between visits.

D. Role of Machine Learning

The relationship between environmental sensor readings and eventual crop yield is highly nonlinear: the effect of a given soil-moisture deficit, for example, depends on the crop's growth stage, the concurrent temperature, and the cumulative water balance over preceding weeks. Classical statistical models such as simple linear regression struggle to represent these interactions and thresholds. Machine learning algorithms, by contrast, are designed to learn nonlinear mappings directly from data and can accommodate high-dimensional, correlated inputs without requiring the relationships to be specified in advance. This makes machine learning a natural complement to sensor-based agricultural monitoring, where the goal is to convert a stream of raw environmental readings into an actionable yield estimate.

E. Research Problem

Despite growing interest in both agricultural IoT and machine learning individually, there remains a need for a clearly specified, end-to-end framework that describes how raw, multi-parameter sensor data should be acquired, cleaned, transformed into predictive features, and used to train and compare regression models specifically for crop yield estimation. Many existing studies focus narrowly on a single sensor type, a single algorithm, or a purely remote-sensing perspective, leaving a gap for a framework that treats ground-based sensor data as the primary input and systematically compares multiple learning algorithms under a common preprocessing and evaluation pipeline.

F. Research Objectives

1. To review the role of IoT sensor data and machine learning techniques in crop yield prediction and identify gaps in the existing literature.
2. To design a structured, end-to-end pipeline for acquiring, storing, and preprocessing multi-parameter agricultural sensor data.

3. To define a feature-engineering scheme that converts raw sensor streams into predictive features suitable for regression modelling.
4. To describe and conceptually compare six representative machine learning algorithms for yield prediction from sensor data.
5. To specify an appropriate set of evaluation metrics for assessing regression-based yield prediction models.
6. To outline a decision-support architecture through which yield predictions can inform irrigation, fertilization, and harvest planning.
7. To identify the practical limitations and future research directions relevant to deploying such a framework in real agricultural settings.

G. Contributions of the Study

- A consolidated, sensor-parameter-driven framework for machine-learning-based crop yield prediction, covering the full pipeline from field sensing to farmer-facing decision support.
- A structured comparison of six regression-based machine learning algorithms with respect to their suitability for noisy, heterogeneous agricultural sensor data.
- A literature comparison table summarizing methods, data sources, and limitations of prior sensor- and remote-sensing-based yield prediction studies.
- An explicit, clearly labelled experimental and evaluation design that can be populated with real data in future implementation work, avoiding unsupported claims of empirical accuracy.

II. LITERATURE REVIEW

This section reviews prior work relevant to machine learning in precision agriculture, sensor-based crop monitoring, and regression- and deep-learning-based yield prediction, and identifies the specific gap that the proposed framework addresses.

A. Machine Learning in Precision Agriculture and Yield Prediction

Van Klompenburg et al. [1] conducted a systematic literature review of machine learning applications in crop yield prediction, finding that ensemble methods and neural networks were increasingly favored over simple linear models as the number of predictor variables grew, and that feature selection played a decisive role in model accuracy. Chlingaryan et al. [2] reviewed machine learning approaches specifically for crop yield prediction and nitrogen status estimation in precision agriculture, noting that most existing work combined remote-sensing vegetation indices with limited ground-truth sensor data rather than dense, continuous in-field sensing. Rashid et al. [3] presented a comprehensive review of machine-learning-based crop yield

prediction with particular emphasis on palm oil, cataloguing the input variables, algorithms, and evaluation metrics used across dozens of studies and observing that Random Forest and ensemble tree methods were consistently among the top performers. Bali and Singla [4] surveyed emerging trends in machine learning for crop yield prediction and the environmental, soil, and management factors that most strongly influence model outcomes, emphasizing the growing role of deep learning as datasets grow larger.

B. Ensemble and Tree-Based Regression Methods

Breiman [5] introduced the Random Forest algorithm as an ensemble of decorrelated decision trees trained on bootstrap-resampled data, a method that has since become one of the most widely used regressors in agricultural yield studies because of its robustness to noisy and correlated features. Jeong et al. [6] applied Random Forests to global and regional crop yield prediction for wheat, maize, and potato, and found that Random Forest models substantially outperformed multiple linear regression benchmarks, with root-mean-square errors between 6% and 14% of observed yield compared with 14% to 49% for the linear baseline. Comparative work reviewed by Discover Agriculture [7] similarly reported that Random Forest and Extreme Gradient Boosting (XGBoost) models achieved the highest coefficients of determination among several algorithms tested on meteorological and yield data for potatoes, maize, and cotton, reinforcing the suitability of tree-based ensembles for this problem domain.

C. Deep Learning and Remote-Sensing-Based Prediction

Nevavuori et al. [8], [9] examined the use of deep convolutional neural networks on multitemporal unmanned-aerial-vehicle (UAV) imagery for crop yield prediction, showing that spatio-temporal deep learning models could extract yield-relevant patterns directly from imagery without hand-crafted features. Shook et al. [10] integrated genotype and weather variables using deep learning to predict crop yield, illustrating how neural architectures can combine heterogeneous data sources. Ju et al. [11] compared several machine learning models for county-level crop yield prediction using MODIS-derived vegetation indices together with weather data, concluding that model choice and feature combination both had a marked effect on prediction accuracy. Hara et al. [12] used artificial neural networks with remote-sensing-derived independent variables for yield prediction and discussed systematic approaches to variable selection. Sun et al. [13], Cao et al. [14], and Qiao et al. [15] each applied deep spatial-temporal architectures — including multilevel networks and recurrent 3-D convolutional networks — to satellite-derived, multi-source

data for county- or regional-level corn and wheat yield estimation, extending the deep-learning approach from single-image inputs to full growing-season time series.

D. Sensor- and Soil-Data-Driven Approaches

A number of studies have specifically examined ground-based sensor networks rather than remote sensing as the primary data source. Bhojani and Bhatt [16] proposed new activation functions for neural networks applied to wheat crop yield prediction from field measurements, reporting improved convergence compared with standard activation functions. Kung et al. [17] used an ensemble neural network method to improve the accuracy of agricultural data analysis. Haghverdi et al. [18] predicted cotton lint yield from phenology-derived crop indices using artificial neural networks, demonstrating the value of combining phenological and environmental descriptors. More recent IoT-oriented work has begun to combine sensor acquisition with on-device or cloud-based learning: Kumar et al. [19] proposed a deep-learning approach to predict crop yield directly from IoT sensor data, and Apat et al. [20] described an IoT-assisted crop-monitoring framework using machine learning algorithms for smart farming, both of which are close in spirit to the framework proposed in this paper but differ in the specific sensor set, preprocessing pipeline, and set of algorithms compared.

E. Broader Reviews and Supporting Literature

Liakos et al. [21] provided an early and widely cited review of machine learning applications across the whole of agriculture, including yield prediction, disease detection, and livestock management, and identified data availability and sensor heterogeneity as recurring obstacles. Tian et al. [22] and Wang et al. [23] combined Long Short-Term Memory (LSTM) and hybrid convolutional-recurrent architectures with remote-sensing and meteorological data to improve wheat yield estimates, illustrating the continuing shift toward temporally aware deep-learning models. Yadav and Geli [24] used a century-scale climate and remote-sensing dataset to predict crop yield for New Mexico, demonstrating the value of long historical baselines. Kheir et al. [25] developed an automated machine-learning pipeline that fused remote sensing, soil, and weather data for fast, robust yield prediction, an approach conceptually aligned with the multi-source fusion adopted in the framework proposed here.

F. Literature Comparison Table

Table I summarizes representative studies discussed above, organized by method, data source, target crop or application, and principal limitation.

Author/Year	Method	Data Used	Crop/Application	Key Findings	Limitation
Van Klompenburg et al. [1], 2020	Systematic review (multiple ML)	Literature synthesis	General crop yield	Ensemble/NN methods trend upward; feature selection is decisive	No new empirical data; review-level only
Chlingaryan et al. [2], 2018	Review (ML for N-status & yield)	Remote sensing + limited ground sensors	Precision agriculture, general	Ground-truth sensor data underused relative to remote sensing	Sparse in-field sensor coverage in reviewed work
Rashid et al. [3], 2021	Review (RF, SVM, ANN, etc.)	Literature synthesis	Palm oil emphasis, general crops	RF and ensembles consistently top performers	Crop-specific bias toward palm oil studies
Bali & Singla [4], 2022	Survey of ML trends	Literature synthesis	General crop yield	Deep learning role growing with data volume	Survey-level; no unified experimental protocol
Breiman [5], 2001	Random Forest (methodology)	General ensemble learning theory	Not crop-specific	Bagged trees reduce variance vs. single trees	Original method paper; not agriculture-specific
Jeong et al. [6], 2016	Random Forest vs. MLR	Climate + biophysical variables	Wheat, maize, potato	RF outperformed linear regression at global/regional scale	Coarse spatial resolution; limited sensor granularity
Discover Agriculture study [7], 2025	RF, XGBoost, GBM, CNN comparison	Meteorological + yield records	Potato, maize, cotton	RF/XGBoost gave highest R ² among models tested	Meteorological data only; no soil sensor input
Nevavuori et al. [8],[9], 2019–2020	Deep CNN, spatio-temporal DL	UAV multitemporal imagery	General field crops	DL extracts yield patterns without hand-crafted features	Requires UAV infrastructure; costly to scale
Shook et al. [10], 2021	Deep learning (genotype+weather)	Genotype and weather records	Multi-environment trials	Combines heterogeneous data types effectively	Requires genotype data, often unavailable to farmers
Ju et al. [11], 2021	ML model comparison	MODIS vegetation indices + weather	County-level yield	Model and feature choice both matter substantially	Remote-sensing resolution limits field-level precision
Hara et al. [12], 2021	Artificial Neural Network	Remote sensing variables	General crops	Systematic variable selection improves ANN accuracy	Variable selection process is data-set specific
Bhojani & Bhatt [16], 2020	ANN with novel activation functions	Field-measured crop data	Wheat	New activation functions improved convergence	Single-crop, single-region validation
Haghverdi et al. [18], 2018	Artificial Neural Network	Phenology-derived crop indices	Cotton	Phenological indices useful predictors of lint yield	Indices require imagery-derived preprocessing
Kumar et al. [19], 2023	Deep learning on IoT sensor data	IoT soil/weather sensors	General field crops	Direct sensor-to-yield deep learning pipeline	Limited description of preprocessing/evaluation protocol
Apat et al. [20], 2022	ML-based IoT crop monitoring	IoT sensor network	Smart farming, general	Demonstrates feasibility of IoT-ML integration	Focused on monitoring; yield-prediction evaluation limited
Liakos et al. [21], 2018	Review (ML across agriculture)	Literature synthesis	General agriculture	Identifies sensor heterogeneity as a recurring obstacle	Broad scope dilutes yield-specific depth

TABLE I. Literature Comparison Summary

G. Research Gap

The reviewed literature shows two broad clusters: reviews and comparative studies that synthesize algorithmic performance without a unifying sensor-driven pipeline, and remote-sensing- or imagery-centric deep-learning studies that operate at coarse spatial or temporal resolution. Relatively few studies specify a complete, reproducible pipeline that begins with a defined set of ground-based, multi-parameter IoT sensors, applies an explicit preprocessing and feature-engineering sequence, and then systematically compares classical and ensemble regression algorithms under common evaluation metrics. This gap motivates the framework proposed in Section III, which is deliberately organized around sensor-level inputs rather than remote-sensing imagery alone, while remaining compatible with the integration of remote-sensing features in future extensions.

III. PROPOSED SYSTEM / RESEARCH METHODOLOGY

This section describes the proposed methodology for machine-learning-based crop yield prediction using sensor data, organized around the system architecture, sensor data collection, dataset description, preprocessing, feature engineering, and the candidate machine learning models.

A. System Architecture

The proposed pipeline follows the sequence: Agricultural Field → Sensors → Data Acquisition → IoT Gateway → Data Storage → Preprocessing → Feature Engineering → ML Model → Yield Prediction → Farmer Decision Support, illustrated in Fig. 1. Sensors deployed across the field continuously measure soil and atmospheric parameters. Readings are collected at the sensor node and transmitted, typically over a low-power wireless protocol, to a local IoT gateway, which performs initial buffering and forwards the data to cloud or on-premise storage. From storage, the raw readings pass through a preprocessing and feature-engineering stage before being supplied to a trained machine learning model, whose output — a predicted yield value — is finally surfaced to the farmer or agricultural decision-maker through a dashboard or mobile interface.



Fig. 1. Overall System Architecture

B. Sensor Data Collection

Table II lists the sensor parameters considered by the proposed framework, together with representative sensor types and their role in yield prediction. Fig. 2 illustrates how these sensors are distributed across the field and aggregated

at an IoT gateway before transmission to storage. It should be noted that specific sensor hardware is described here as representative of common low-cost agricultural sensing equipment; no physical sensor deployment was carried out for this study, and the parameter list is intended to define the input space that the proposed framework is designed to accept.

Sensor Parameter	Sensor Type / Example	Unit	Purpose in Prediction
Soil Moisture	Capacitive/resistive soil moisture probe	%VWC	Indicates water availability and irrigation need
Soil Temperature	Digital soil temperature probe (e.g., DS18B20-type)	°C	Affects germination, root activity, and nutrient uptake
Air Temperature	Thermistor/digital temperature sensor	°C	Drives growth rate and evapotranspiration
Relative Humidity	Capacitive humidity sensor	%RH	Influences disease pressure and transpiration
Rainfall	Tipping-bucket rain gauge	mm	Direct water input and runoff risk indicator
Soil pH	Ion-selective pH probe	pH units	Governs nutrient solubility and uptake efficiency

Sensor Parameter	Sensor Type / Example	Unit	Purpose in Prediction
Electrical Conductivity (EC)	Soil EC probe	dS/m	Proxy for soil salinity and dissolved nutrients
NPK Levels	Nutrient (N, P, K) sensor/soil test kit	mg/kg	Directly linked to plant nutrition and growth potential
Light Intensity	LDR / lux sensor	lux	Relates to photosynthetically active radiation available

TABLE II. Sensor Parameters and Their Agricultural Significance

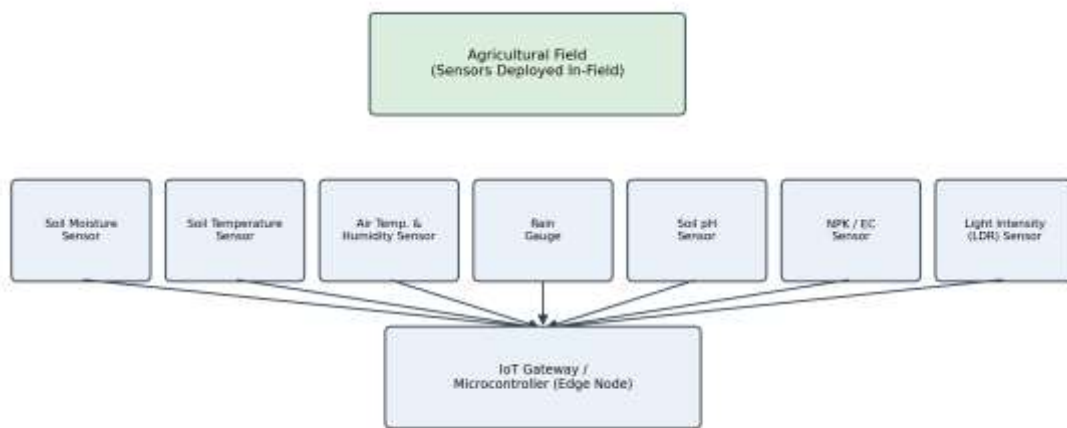


Fig. 2. Sensor Data Collection Framework

C. Dataset Description

The proposed framework assumes a tabular time-series dataset in which each record corresponds to a sensor reading (or a daily/weekly aggregate of readings) for a specific field plot, together with the eventual harvested yield for that plot at the end of the growing season. Table III summarizes the intended dataset structure. Where a public dataset is used to

instantiate the framework, an appropriate source would be an openly available agricultural sensor or crop-recommendation dataset, such as those hosted on Kaggle or provided by national agricultural research institutions; the specific dataset size and provenance should be documented explicitly at the point of implementation rather than assumed in advance, and no specific dataset size is claimed in this proposal.

Attribute	Description
Input variables	Soil moisture, soil temperature, air temperature, humidity, rainfall, pH, EC, NPK, light intensity
Target variable	Harvested yield per unit area (e.g., kg/hectare or tonnes/hectare)
Sampling frequency	Hourly to daily sensor readings, aggregated to daily/weekly summaries
Data records	To be determined at implementation, based on field size, season length, and sensor count
Crop type	Configurable; framework is crop-agnostic but should be calibrated per crop
Growth period	Full season, from sowing/planting to harvest
Location	Field-plot level, ideally with geo-tagging for regional generalization studies
Yield measurement	Post-harvest measured yield, recorded per plot or per sensor cluster

TABLE III. Dataset Attributes

D. Data Preprocessing

Raw sensor data collected under field conditions is rarely suitable for direct use in model training. The proposed preprocessing sequence, shown in Fig. 3, addresses eight concerns:

- **Missing-value handling:** gaps arising from sensor downtime or transmission failure are addressed through interpolation (for short gaps) or model-based imputation (for longer gaps), rather than naive deletion, to preserve the continuity of the time series.
- **Duplicate removal:** repeated transmissions or logging errors are identified through timestamp and sensor-ID matching and removed.
- **Outlier detection:** implausible readings (e.g., soil moisture above 100% or negative rainfall) are flagged using range checks and statistical methods such as the interquartile range (IQR) rule.
- **Noise filtering:** small-magnitude sensor noise is reduced using moving-average or Savitzky–Golay smoothing over short time windows.
- **Normalization/standardization:** features are rescaled, typically via z-score standardization, $x' = (x - \mu) / \sigma$, so that variables measured on different scales (e.g., pH versus rainfall in millimetres) contribute comparably during training.
- **Feature encoding:** categorical attributes such as crop type or soil class are converted to numeric form via one-hot or ordinal encoding as appropriate.
- **Data synchronization:** readings from sensors with different sampling rates are aligned to a common time index (e.g., daily) through aggregation or resampling.
- **Train-test splitting:** the cleaned dataset is partitioned, commonly 70–80% for training and 20–30% for testing, with care taken to split by season or plot rather than randomly within a single time series, to avoid information leakage between temporally adjacent records.



Fig. 3. Data Preprocessing Pipeline

E. Feature Engineering

Beyond cleaning, raw sensor streams are transformed into higher-level features that are more directly predictive of yield. Examples include: average and cumulative soil moisture over the growing season; maximum, minimum, and mean temperature per growth stage; temperature variation (diurnal range); cumulative rainfall and dry-spell counts; growing-degree-day indicators derived from temperature thresholds; humidity averages over critical flowering or fruiting periods; nutrient ratios such as N:P:K balance; and soil-moisture trend features capturing the rate of change rather than absolute level. These engineered features are intended to encode agronomically meaningful patterns — such as sustained water stress or a favorable nutrient balance — that raw instantaneous readings alone do not capture.

F. Machine Learning Models

Six representative regression algorithms are considered, spanning simple linear models, tree-based methods, kernel methods, and neural networks. Table IV summarizes their comparative characteristics.

1) Linear Regression

Linear Regression models yield as a weighted linear combination of input features, $y = \beta_0 + \sum \beta_i x_i + \epsilon$, with coefficients estimated by minimizing squared error. It is simple, fast to train, and highly interpretable, making it a natural baseline, but it cannot capture the nonlinear thresholds and interactions that typically govern the relationship between sensor readings and yield.

2) Decision Tree Regression

A Decision Tree recursively partitions the feature space into regions of similar target value, selecting splits that minimize within-region variance. Decision trees can capture nonlinear relationships and variable interactions without requiring a specified functional form, and are easy to interpret, but a single tree is prone to overfitting and can be unstable to small changes in the training data.

3) Random Forest Regression

Random Forest [5] trains an ensemble of decision trees on bootstrap-resampled subsets of the data and a random subset of features at each split, then averages the individual tree predictions. This bagging approach reduces the variance associated with a single tree while retaining the ability to model nonlinear relationships, and it has repeatedly been found to perform strongly on agricultural sensor and climate

data [6], [7]. Its main disadvantages are reduced interpretability relative to a single tree and comparatively higher computational and memory cost.

4) Support Vector Regression (SVR)

Support Vector Regression fits a function that lies within an ϵ -insensitive margin of the training targets while penalizing points outside that margin, and can use kernel functions (e.g., radial basis function) to model nonlinear relationships. SVR can perform well on moderately sized, high-dimensional datasets and is relatively robust to outliers, but its performance is sensitive to kernel and hyperparameter choice, and training time scales poorly with very large datasets.

5) Gradient Boosting / XGBoost

Gradient Boosting builds an ensemble of shallow trees sequentially, with each new tree trained to correct the residual errors of the ensemble so far. XGBoost is a widely

used, regularized, and computationally efficient implementation of this idea. Gradient boosting methods have shown strong performance in several yield-prediction comparisons [7], [15], often matching or exceeding Random Forest, but require careful tuning of learning rate, tree depth, and the number of boosting rounds to avoid overfitting.

6) Artificial Neural Network (ANN)

A feed-forward Artificial Neural Network consists of layers of interconnected nodes that apply learned weights and nonlinear activation functions to model complex relationships between sensor inputs and yield. ANNs can, in principle, approximate arbitrarily complex nonlinear functions and have been used successfully for yield prediction in several studies [16], [18], but they typically require larger training datasets, more careful regularization to avoid overfitting, and offer less direct interpretability than tree-based methods.

Model	Advantages	Disadvantages	Suitability
Linear Regression	Simple, fast, interpretable	Cannot capture nonlinearity	Baseline model only
Decision Tree	Captures nonlinearity, interpretable	Prone to overfitting, unstable	Exploratory analysis
Random Forest	Robust to noise, handles nonlinearity well	Less interpretable, higher cost	Primary candidate for tabular sensor data
Support Vector Regression	Robust to outliers, effective in high dimensions	Sensitive to kernel/hyperparameters, scales poorly	Moderate-sized, well-scaled datasets
Gradient Boosting / XGBoost	High accuracy, handles nonlinearity and interactions	Requires careful tuning, risk of overfitting	Strong candidate when tuned carefully
Artificial Neural Network	Models complex nonlinear relationships	Needs larger data, less interpretable	Larger, richer sensor datasets

TABLE IV. Comparison of Machine Learning Algorithms

IV. PROPOSED MACHINE LEARNING FRAMEWORK

Building on the architecture and models described in Section III, this section specifies the end-to-end workflow as a step-by-step algorithm, illustrated in Fig. 4.

8. Initialize the sensor network across the target field, covering representative soil and micro-climate zones.
9. Continuously collect environmental and soil data (moisture, temperature, humidity, rainfall, pH, EC, NPK, light).
10. Transmit and store sensor readings in a time-indexed database via the IoT gateway.
11. Clean the dataset: remove duplicates and clearly invalid readings.
12. Handle missing values through interpolation or model-based imputation.
13. Detect and treat outliers using range checks and statistical rules (e.g., IQR).

14. Perform feature engineering to derive season-level and stage-level predictive features.
15. Split the dataset into training and testing subsets, partitioned by plot or season.
16. Train candidate ML models (Linear Regression, Decision Tree, Random Forest, SVR, Gradient Boosting/XGBoost, ANN).
17. Tune model hyperparameters using cross-validation on the training subset.
18. Generate yield predictions on the held-out test subset.
19. Evaluate predictions using MAE, MSE, RMSE, and R^2 (Section VI).
20. Select the best-performing model based on the evaluation metrics and computational cost.
21. Provide the selected model's predictions to the farmer or agricultural decision-support system.

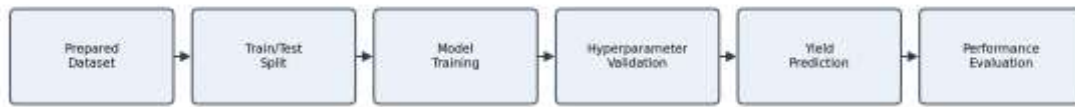


Fig. 4. Machine Learning Prediction Workflow

A simplified pseudocode representation of the training and evaluation loop is as follows:

```

for model in [LinearRegression, DecisionTree,
RandomForest, SVR, GradientBoosting, ANN]:
    fit model on X_train, y_train
    y_pred = model.predict(X_test)
    compute MAE, MSE, RMSE, R2(y_test, y_pred)
    record results
select model with best validation performance
  
```

V. EXPERIMENTAL SETUP

This section describes how the proposed framework can be experimentally evaluated once field or public sensor data is obtained. As no field trial was conducted for this proposal, the setup below is presented as the intended implementation configuration rather than a record of hardware or software actually deployed.

- Hardware requirements: low-power soil and atmospheric sensor nodes, an IoT gateway (e.g., a microcontroller with wireless connectivity), and a workstation or cloud instance for model training.
- Software requirements: Python 3.x as the primary development language.
- ML libraries: Scikit-learn for classical/ensemble models; TensorFlow or PyTorch where an ANN or deep-learning variant is used.
- Supporting libraries: Pandas and NumPy for data handling; Matplotlib for visualization.
- Development environment: Jupyter Notebook for iterative experimentation.
- Data platform: a time-series-capable database or IoT platform for sensor data storage, where the framework is deployed with live sensors.
- Preprocessing process: as described in Section III-D, applied consistently across all candidate models.
- Training procedure: k-fold cross-validation on the training partition, with hyperparameters selected via grid or randomized search.
- Validation procedure: hold-out evaluation on a temporally or spatially separated test partition to reduce the risk of information leakage.

VI. PERFORMANCE EVALUATION

Because crop yield prediction is a regression problem, the following four standard metrics are proposed for evaluating and comparing model performance.

Mean Absolute Error (MAE)

$MAE = (1/n) \sum |y_i - \hat{y}_i|$, the average absolute difference between predicted and actual yield. MAE is easy to interpret in the original yield units and is not overly sensitive to a small number of large errors.

Mean Squared Error (MSE)

$MSE = (1/n) \sum (y_i - \hat{y}_i)^2$, the average of squared errors. Because errors are squared before averaging, MSE penalizes large deviations more heavily than MAE, which is useful when large prediction errors are especially costly for downstream decisions such as fertilizer or irrigation planning.

Root Mean Squared Error (RMSE)

$RMSE = \sqrt{MSE}$, which returns the error to the original yield units while retaining MSE's sensitivity to large errors, making it one of the most commonly reported metrics in the yield-prediction literature [6].

Coefficient of Determination (R^2)

$R^2 = 1 - [\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2]$, the proportion of variance in observed yield that is explained by the model, where $\bar{y}$ (mean of y) is the mean observed yield. An R^2 close to 1 indicates that the model explains most of the yield variability, while values near 0 indicate performance no better than predicting the mean.

Each metric offers a different lens on model quality: MAE and RMSE express error magnitude in interpretable yield units (important for communicating results to farmers), MSE and RMSE emphasize the cost of large errors (important where a single badly mispredicted yield could lead to a poor irrigation or fertilizer decision), and R^2 expresses overall explanatory power, useful for comparing models across datasets of different scale.

VII. RESULTS AND ANALYSIS

No field sensor dataset was collected or analyzed for this study, and no crop was physically monitored. Consequently, this section presents a proposed, illustrative results framework rather than genuine experimental findings. After implementation and evaluation on a real dataset, the proposed framework can be assessed using the table and figures below populated with actual computed values; the entries shown here are placeholders that describe the

intended structure of the analysis and must not be interpreted as measured accuracy.

Model	MAE	MSE	RMSE	R ²	Training Time
Linear Regression	—	—	—	—	—
Decision Tree	—	—	—	—	—
Random Forest	—	—	—	—	—
Support Vector Regression	—	—	—	—	—
Gradient Boosting / XGBoost	—	—	—	—	—
Artificial Neural Network	—	—	—	—	—

TABLE V. Model Evaluation Metrics — To Be Populated From Actual Experimental Data (Not Measured In This Study)

To illustrate what the evaluation outputs would look like once genuine data is available, Figs. 5–7 show representative, clearly labelled illustrative visualizations rather than results derived from this study. Fig. 5 shows an illustrative feature-importance ranking of the type a tree-based model would typically produce, indicating which sensor-derived features tend to carry the most predictive weight in comparable published studies [3], [6]. Fig. 6 shows an illustrative actual-versus-predicted scatter plot of the kind

used to visually assess calibration, with points expected to cluster around the diagonal for a well-fitted model. Fig. 7 shows an illustrative comparison of R² across the six candidate algorithms, consistent in relative ordering with prior comparative studies that found Random Forest and Gradient Boosting to outperform simple linear regression [6], [7], though the specific values shown are for illustration only and are not derived from an executed experiment in this work.

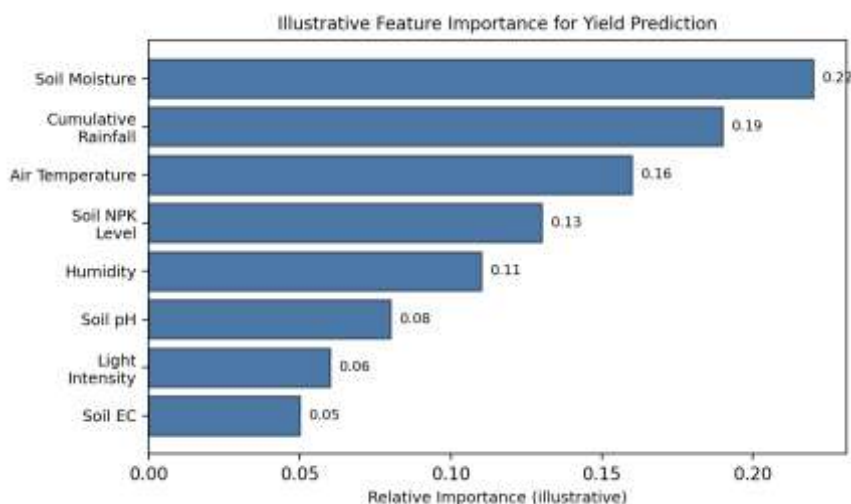


Fig. 5. Illustrative Feature Importance Analysis (not measured in this study)

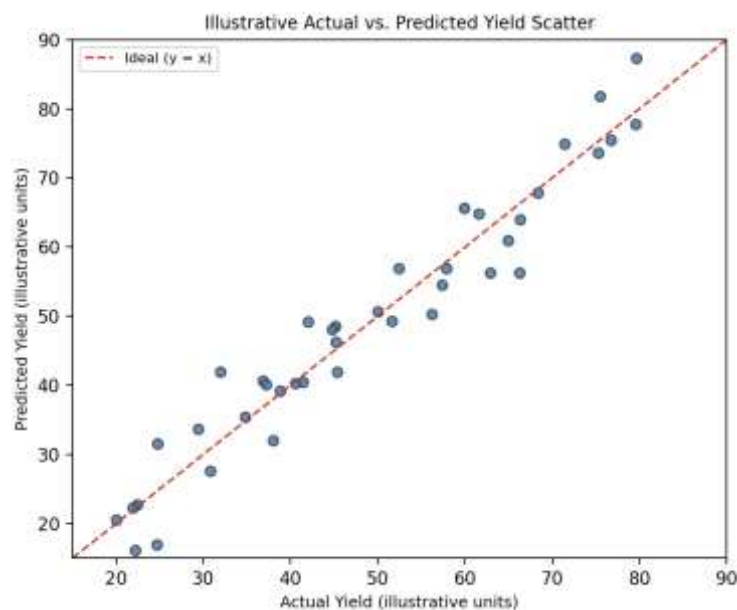


Fig. 6. Illustrative Actual vs. Predicted Yield (not measured in this study)

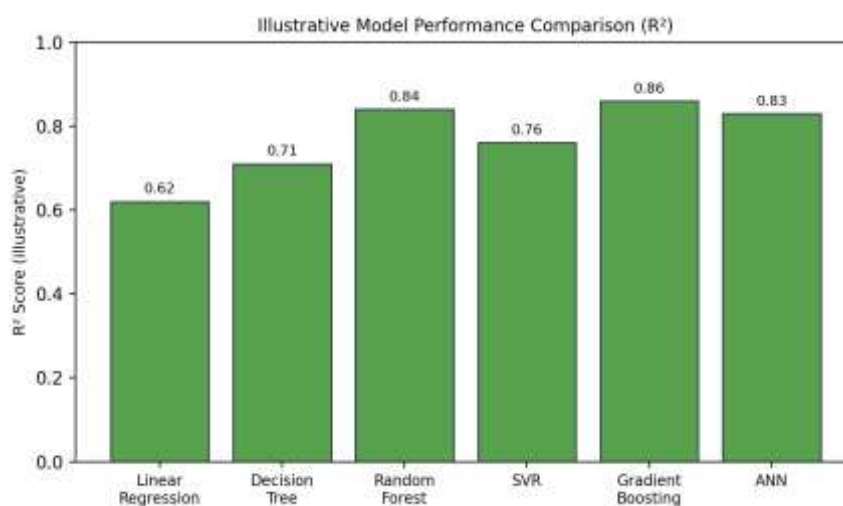


Fig. 7. Illustrative Model Performance Comparison (not measured in this study)

Beyond the specific illustrative values, the qualitative pattern expected — and consistent with the reviewed literature — is that ensemble tree-based methods (Random Forest, Gradient Boosting) and, given sufficient data, Artificial Neural

Networks, should outperform simple Linear Regression on sensor-derived agricultural data because of their capacity to model nonlinear thresholds such as the point at which soil moisture becomes limiting for growth.

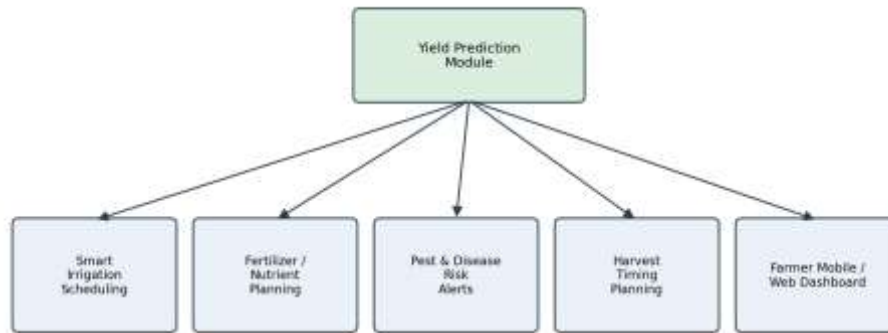


Fig. 8. Proposed Smart Farming Decision-Support Architecture

Fig. 8 shows how the yield-prediction output of the proposed framework is intended to feed into downstream decision-support functions — irrigation scheduling, fertilizer planning, pest and disease risk alerting, harvest-timing planning, and a farmer-facing dashboard — rather than terminating at a single numerical forecast.

VIII. DISCUSSION

Sensor-derived data improves crop yield prediction primarily by replacing infrequent, subjective observation with continuous, quantitative measurement of the variables that most directly govern plant growth. Continuous monitoring makes it possible to detect short-lived stress events — a brief dry spell, a nutrient dip, an unseasonal heat event — that would be invisible to periodic manual inspection but that can materially affect final yield if they coincide with a sensitive growth stage. The relationship between these environmental parameters and yield is rarely linear: soil moisture, for instance, may have little effect on yield above a threshold but a sharp effect below it, and machine learning models are far better suited than traditional linear statistics to represent such thresholds and interactions.

Compared with traditional prediction, machine-learning-based approaches trade some interpretability for accuracy, particularly with ensemble and neural-network models. Simpler models such as Linear Regression remain valuable as an interpretable baseline and as a sanity check on more complex models. Interpretability challenges become more pressing as models grow more complex; techniques such as feature-importance ranking (Fig. 5) and partial-dependence analysis can partially address this by indicating which sensor parameters most influence a given prediction, which matters for farmer trust and for actionable recommendations.

Several practical concerns affect real-world deployment. Sensor reliability is not guaranteed: probes can drift out of calibration, fail in wet or high-temperature conditions, or

transmit corrupted readings, and the preprocessing pipeline in Section III-D must be robust to these failure modes rather than assuming clean data. Missing data is common in field deployments due to connectivity gaps, and imputation strategies must be chosen carefully to avoid introducing bias. Model generalization across different crops and geographical conditions is not automatic: a model trained on sensor data from one region's soil type and climate may not transfer well to another, and recalibration or region-specific retraining is generally necessary. Scalability and cost considerations — the price of sensors, connectivity, and cloud infrastructure relative to the value of the crop — will determine whether such a system is economically viable for smallholder farmers versus large commercial operations. Finally, integration with broader precision-agriculture practices, such as variable-rate irrigation and fertilization equipment, is what ultimately converts an accurate prediction into an actionable and profitable outcome.

IX. REAL-WORLD APPLICATIONS

22. Smart irrigation: scheduling water application based on predicted soil-moisture deficits rather than a fixed calendar.
23. Fertilizer optimization: tailoring nutrient application to predicted yield potential and measured soil nutrient status.
24. Crop planning: informing which crop or variety to plant based on historical sensor-yield relationships for a given plot.
25. Harvest planning: anticipating harvest timing and expected volume to coordinate labor and storage logistics.
26. Early identification of yield decline: flagging fields where sensor trends suggest a developing stress condition before visible symptoms appear.

27. Farm resource optimization: prioritizing limited water, labor, or input budgets across multiple fields based on predicted marginal benefit.

28. Precision agriculture integration: feeding predictions into variable-rate application equipment for site-specific management.

29. Agricultural decision-support systems: providing farmers and advisors with a consolidated, data-driven view of expected outcomes.

30. Greenhouse management: adapting the same sensor-to-prediction pipeline to controlled-environment agriculture, where sensor density is typically even higher.

31. Government and agricultural planning: aggregating plot-level predictions to support regional food-security and supply planning.

X. LIMITATIONS

Table VI summarizes the principal limitations of the proposed framework.

Limitation	Description
Sensor calibration	Drift over time can bias readings and downstream predictions if not periodically checked
Sensor failure	Hardware faults or power loss can create data gaps or corrupted readings
Missing data	Connectivity or transmission failures reduce the completeness of the training data
Dataset size	Field-collected agricultural datasets are often small relative to the number of predictive features
Regional variation	Models trained in one region or soil type may not generalize to another without recalibration
Crop-specific behavior	Sensor-yield relationships differ across crop species and cultivars
Weather uncertainty	Unpredictable extreme weather events are difficult to anticipate from historical patterns alone
Soil heterogeneity	Within-field variability in soil type can reduce the representativeness of a limited sensor network
Model bias	Models can inherit biases present in historically observed conditions and underperform on novel conditions
Deployment cost	Sensor hardware, connectivity, and maintenance costs may be prohibitive for smallholder farmers
Connectivity requirements	Reliable internet or wireless connectivity is not universally available in rural areas

TABLE VI. Limitations of the Proposed System

These limitations mean the proposed system should not be regarded as universally accurate; its performance will depend on the quality, density, and duration of sensor deployment, and on careful, region- and crop-specific validation before being relied upon for high-stakes decisions.

XI. FUTURE WORK

- Real-time edge AI: running lightweight trained models directly on IoT gateways to reduce latency and connectivity dependence.
- Federated learning: training shared models across multiple farms without centralizing raw sensor data, addressing privacy and bandwidth concerns.

- Explainable AI: applying interpretability techniques (e.g., SHAP, LIME) to build farmer trust in model recommendations.
- Deep learning: exploring recurrent and transformer-based architectures for richer temporal modelling of sensor sequences.
- Satellite imagery integration: fusing ground-sensor data with satellite-derived vegetation indices for broader spatial coverage.
- Drone imagery: incorporating UAV-based imagery for higher-resolution, in-season crop-health assessment.
- Weather API integration: supplementing on-site sensors with forecast data to anticipate upcoming stress conditions.
- Digital twins for agriculture: maintaining a continuously updated virtual model of the field for simulation-based decision support.

- Multicrop prediction: extending the framework to jointly model multiple crops grown in rotation or intercropped.
- Mobile applications: delivering predictions and recommendations directly to farmers through accessible mobile interfaces.
- Low-cost IoT deployment: developing cheaper sensor and connectivity options suitable for smallholder adoption.
- Automated irrigation: closing the loop by directly actuating irrigation systems based on model output, subject to safety checks.
- Reinforcement learning for farm decisions: exploring sequential decision-making frameworks that adapt management actions over a growing season.

XII. CONCLUSION

This paper has proposed a machine-learning-based framework for crop yield prediction driven by multi-parameter agricultural sensor data, addressing the limitations of manual, experience-based yield estimation through continuous, automated environmental monitoring combined with data-driven modelling. The framework specifies a complete pipeline from sensor deployment through preprocessing, feature engineering, model training, and evaluation, and conceptually compares six regression algorithms — Linear Regression, Decision Tree, Random Forest, Support Vector Regression, Gradient Boosting/XGBoost, and Artificial Neural Networks — for their suitability to this problem. Because no field dataset was collected for this study, the work is presented explicitly as a proposed framework and experimental design rather than a report of measured results, and all illustrative figures and tables are labelled accordingly. If implemented with genuine sensor data, the framework is expected to support practical benefits for farmers and agricultural planners, including more timely irrigation and fertilization decisions and earlier identification of yield risk. Realizing this potential in practice will require addressing the limitations discussed in Section X, particularly sensor reliability, cost, and regional generalization, and pursuing the future directions outlined in Section XI, from edge AI to satellite-data fusion, to move the framework from a conceptual proposal toward a field-validated decision-support tool.

ACKNOWLEDGMENT

The author gratefully acknowledges the guidance and support of the faculty guide and the Department of Computer Science and Engineering, Kings College of Engineering, Punalkulam, Tamil Nadu, India, in the preparation of this work. Appreciation is also extended to the institution for providing the academic environment and resources that made this research possible.

REFERENCES

- [1] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
- [2] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and Electronics in Agriculture*, vol. 151, pp. 61–69, 2018, doi: 10.1016/j.compag.2018.05.012.
- [3] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," *IEEE Access*, vol. 9, pp. 63406–63439, 2021, doi: 10.1109/ACCESS.2021.3075159.
- [4] N. Bali and A. Singla, "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey," *Archives of Computational Methods in Engineering*, vol. 29, no. 1, pp. 95–112, 2022, doi: 10.1007/s11831-021-09569-8.
- [5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [6] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy, and S.-H. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, e0156571, 2016, doi: 10.1371/journal.pone.0156571.
- [7] "A comparative study of machine learning models in predicting crop yield," *Discover Agriculture*, vol. 3, 2025, doi: 10.1007/s44279-025-00335-z.
- [8] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, p. 104859, 2019, doi: 10.1016/j.compag.2019.104859.
- [9] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models," *Remote Sensing*, vol. 12, no. 23, p. 4000, 2020, doi: 10.3390/rs12234000.
- [10] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh, "Crop yield prediction integrating genotype and weather variables using deep learning," *PLoS ONE*, vol. 16, no. 6, e0252402, 2021, doi: 10.1371/journal.pone.0252402.
- [11] S. Ju, H. Lim, J. W. Ma, S. Kim, K. Lee, S. Zhao, and J. Heo, "Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models," *Agricultural and Forest Meteorology*, vol. 307, p. 108530, 2021, doi: 10.1016/j.agrformet.2021.108530.
- [12] P. Hara, M. Piekutowska, and G. Niedbała, "Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data," *Land*, vol. 10, no. 6, p. 609, 2021, doi: 10.3390/land10060609.
- [13] J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, and Y. Shen, "Multilevel deep learning network for county-level corn yield estimation in the U.S. Corn Belt," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5048–5060, 2020, doi: 10.1109/JSTARS.2020.3019046.
- [14] J. Cao, Z. Zhang, F. Tao, L. Zhang, Y. Luo, J. Han, and Z. Li, "Identifying the contributions of multi-source data for winter wheat yield prediction in China," *Remote Sensing*, vol. 12, no. 5, p. 750, 2020, doi: 10.3390/rs12050750.

- [15] M. Qiao, X. He, X. Cheng, P. Li, H. Luo, L. Zhang, and Z. Tian, "Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102436, 2021, doi: 10.1016/j.jag.2021.102436.
- [16] S. H. Bhojani and N. Bhatt, "Wheat crop yield prediction using new activation functions in neural network," *Neural Computing and Applications*, vol. 32, no. 17, pp. 13941–13951, 2020, doi: 10.1007/s00521-020-04797-8.
- [17] H.-Y. Kung, T.-H. Kuo, C.-H. Chen, and P.-Y. Tsai, "Accuracy analysis mechanism for agriculture data using the ensemble neural network method," *Sustainability*, vol. 8, no. 8, p. 735, 2016, doi: 10.3390/su8080735.
- [18] A. Haghverdi, R. A. Washington-Allen, and B. G. Leib, "Prediction of cotton lint yield from phenology of crop indices using artificial neural networks," *Computers and Electronics in Agriculture*, vol. 152, pp. 186–197, 2018, doi: 10.1016/j.compag.2018.07.021.
- [19] V. Kumar, P. Parwekar, and G. Gupta, "Deep learning approach to predict crop yield using IoT sensor data," in *Proc. Int. Conf. ICT for Sustainable Development*, Springer, 2023, pp. 39–46.
- [20] S. K. Apat, J. Mishra, K. S. Raju, and N. Padhy, "IoT-assisted crop monitoring using machine learning algorithms for smart farming," in *Next Generation of Internet of Things: Proc. ICNGIoT 2022*, Springer, 2022, pp. 1–11.
- [21] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018, doi: 10.3390/s18082674.
- [22] H. Tian, P. Wang, K. Tansey, J. Zhang, S. Zhang, and H. Li, "An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China," *Agricultural and Forest Meteorology*, vol. 310, p. 108629, 2021, doi: 10.1016/j.agrformet.2021.108629.
- [23] J. Wang, P. Wang, H. Tian, K. Tansey, J. Liu, and W. Quan, "A deep learning framework combining CNN and GRU for improving wheat yield estimates using time series remotely sensed multi-variables," *Computers and Electronics in Agriculture*, vol. 206, p. 107705, 2023, doi: 10.1016/j.compag.2023.107705.
- [24] K. Yadav and H. M. E. Geli, "Prediction of crop yield for New Mexico based on climate and remote sensing data for the 1920–2019 period," *Land*, vol. 10, no. 12, p. 1389, 2021, doi: 10.3390/land10121389.
- [25] A. M. S. Kheir et al., "Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing, soil, and weather dataset," *Environmental Research Communications*, vol. 6, no. 4, p. 041005, 2024, doi: 10.1088/2515-7620/ad2d02.