International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Hybrid Feature Selection Using Genetic Algorithms and Random Forests for Crop Yield Prediction: A Systematic Optimization Framework for Precision Agriculture

Aryan Kumar Shasana[1], Mrs. Kavisha Yadav[2]

[12]Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, Rajasthan, India

Final Year Undergraduate Research Project - Academic Session 2025–2026

Abstract - *Crop yield prediction is an important problem for global food security and precision agriculture management, macroeconomic supply chain planning, and for farmers to take decisions at the local level. Modern agri-telemetry monitors high dimensional soil chemistry, micro-climate data, satellite-derived spectral vegetation indices, and agronomic management data. These datasets, however, are often burdened with the worry of dimensionality, high level of multicollinearity, redundant parameters and observational noise, all of which negatively impact the generalization capabilities of machine-learning models, the computational efficiency of the models, and the interpretability of the results. A hybrid wrapper-based optimization architecture is proposed in this paper to combine the binary-coded Genetic Algorithm (GA) with a Random Forest (RF) regressor to be used for feature selection in the case of crop yield prediction. The feature masks of the candidates are evolved using evolutionary operators (tournament selection, uniform crossover and adaptive mutation), and are objectively evaluated based on fitness function Root Mean Squared Error (RMSE) using Out-of-Bag (OOB) cross validation penalizing model complexity. Tested with a multi-regional agricultural dataset with 35 features that include soil, climate, and management parameters, the proposed GA-RF framework reduces the dimensionality of the data by 62.8% (choosing 13 optimal features out of 35) with great enhancement in the predictive accuracy. It outperforms Full-Feature Random Forest, Principal Component Analysis (PCA), Recursive Feature Elimination (RFE) and XGBoost with high R2 score (0.948), Root Mean Square Error (RMSE) (0.312 tons/hectare) and Mean Absolute Error (MAE) (0.224 tons/hectare) scores. Moreover, the authors propose an enterprise MLOps deployment architecture, an agronomic interpretability analysis using SHAP, strategies for mitigating concept drift and a SHAP implementation roadmap for real-time deployment on an edge in smart farming scenarios. The project utilizes the expertise of the students in the following areas: Precision Agriculture, Crop Yield Prediction, Genetic Algorithms, Random Forest, Feature Selection, Evolutionary Optimization, MLOps, SHAP Interpretability.*

1. INTRODUCTION AND THEORETICAL MOTIVATION

1.1 Context

Agricultural production is becoming more and more sensitive to the variability of climate, extreme events, soil degradation and the availability of inputs. To meet sustainable food security needs for a nation as large as the globe with a projected population of nearly 10 billion people by 2050, traditional farming methods need to become precision agriculture systems that are based on data. Timely prediction of crop yield before harvest helps the policy makers, agricultural cooperatives, supply chain managers and individual farmers appropriately allocate storage, fix the prices, plan their imports/export and target use of resources (irrigation, nitrogen addition, pesticides etc.). The use of field sensors, multispectral satellite images (e.g., Sentinel-2, Landsat-8), the increasing density of autonomous weather stations, and soil chemistry assays has led to huge quantities of agronomic information over the last few years. These multi-source telematic feeds reflect complex and non-linear interactions between edaphic soil profiles, meteorological dynamics, satellite vegetation indices and agronomic inputs.

1.2 Problem Statement

High-dimensional agricultural datasets create significant challenges for computation and analysis: ˆ Lots of data, lots of teasers – many more features than data points:

- Soil nutrients, ambient temperature, and satellite vegetation indices often are highly correlated with each other. Collinear features increase model variance, and can muddy the split statistics in trees, as well as introduce a higher potential of over fitting.
- Observational Noise and Sensor Drift: Sensor drift and interference by atmosphere in low-cost field sensors and satellite observations add noise to supervised learning algorithms.
- Machine learning models often contain tens or hundreds of raw features and are opaque to the user: this lack of interpretability reduces trust between agronomists and farmers that need to understand the biological mechanisms behind yield predictions.

1.3 Research Objective and Core Contribution

To overcome these problems this paper gives the systematic research investigation at the Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur. The main goal was to design, implement and validate a hybrid architecture of a metaheuristic optimization method-based system that automatically determines the best subset of agronomic features to predict crop yields.

The major contributions of this paper include:

1. Systematic Hybrid Architecture: We design a closed-loop wrapper that combines a binary Genetic Algorithm(GA) with an ensemble Random Forest(RF) regressor, and avoid data leakage in feature subset search by using Out-of-Bag(OOB) cross validation.
2. Penalized Fitness Function Design: We define fitness function specific to the problem that penalizes model complexity (Feature Reduction Rate) and predictive accuracy (RMSE) to drive the evolutionary search towards compact, highly predictive feature masks.
3. Comprehensive Benchmarking: Multi Baseline models are compared with the proposed GA-RF framework using the standardized performance metrics such as RMSE, MAE, R2, training time, and inference latency.
4. Agronomic Explainability via SHAP: Using SHapley Additive exPlanations (SHAP) to explain the biological and physical processes underlying yield prediction by the GA-selected subset of features.
5. Production MLOps Deployment Blueprint: We've created a full enterprise deployment architecture that includes data pipelines, microservice containerization, REST API serving, and automated concept-drift monitoring.

2. LITERATURE REVIEW AND RELATED WORK

2.1 Evolution of Crop Yield Prediction Model

Early crop yield prediction models were based mainly on deterministic agronomic growth simulation models (e.g. DSSAT, APSIM) or historical multivariate linear regression models. Growth models have detailed physiological representations, but are cumbersome to calibrate locally and do not scale easily across large regions geographies. In contrast, the ordinary least square (OLS) regression models are unable to explain complex and non-linear relationships between soil chemistry, micro-climates, and satellite vegetation indices.

Machine learning made a breakthrough with the development of non parametric supervised learning methods that are able to learn non-linear relationships without distributional assumptions. Various machine learning techniques like Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gradient Boosting Machines (GBM), and Random Forests (RF) have shown to perform well in agriculture forecasting tasks. Random Forest has proven itself to be a compact standard in the field of agricultural modeling because of its resistance to outliers, its built-in randomness of its internal aggregation (bootstrap), and its ability to process mixed categorical and continuous features.

2.2 Taxonomy of Feature Selection Methods

Feature selection methods can be broadly categorized into three main types:

1. Filter Methods: Measure the feature importance by its statistical properties separately from the predictive model (e.g., Pearson correlation coefficient, Analysis of Variance [ANOVA], Mutual Information). Computationally fast, filter methods assess features on a one-to-one basis, which means that they do not recognise interaction between multiple features.
2. Embedded Methods: Feature Selection by penalizing weights of features in model training (e.g., Lasso/L1 regularization) or based on the impurity in the tree splits (e.g., Gini importance in Random Forests). Embedded methods are quick, but might favour continuous features with many distinct values.
3. Wrapper Methods: Design a search heuristic to explore the power set of features combinations, and use the predictive model as the evaluator of the combinations (2^d). These wrapper methods can tolerate interactions between features and biases in the models, and can improve predictive accuracy, but can involve more computational searches.

2.3 Metaheuristic Optimization in Agriculture

Some researchers have focused on metaheuristic search algorithms like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Grey Wolf Optimizer (GWO) and Genetic Algorithms (GA) to tackle the efficient exploration of the exponentially large search space (2d) of feature subsets in agriculture. Based on the Darwinian principle of natural selection, Genetic Algorithms are a class of algorithms that are well suited for global search in multi-modal space of combinatorial problems without falling into local minima. Recent research has shown that the use of GAs in conjunction with other features, such as tree-based ensembles, in a variety of applications, such as medical diagnostics and financial risk scoring, can be effective.

This paper proposes an advancement of the recent literature that created the structural review and MLOps deployment paradigms from recent papers (Singh, 2026) and applies the concept of feature optimization with a metaheuristic approach to multi-source agricultural datasets, which links the theoretical aspect of algorithm design with the application in field.

Table 1: Taxonomic Comparison of Feature Selection Paradigms

Paradigm	Search Mechanism	Key Advantages	Key Limitations
Filter	Univariate statistical scores (Pearson, Mutual Info)	Extremely fast ($O(d)$); model agnostic; scalable.	Ignores feature interactions; unable to detect redundancy.
Embedded	Inherent model mechanisms (Lasso L1, Tree Split Gini)	Integrated training ($O(d \cdot N)$); captures local feature importance.	Dependent on model family; biased toward high-cardinality features.
Wrapper (Greedy)	Recursive Feature Elimination (RFE)	Sequential removal based on model weights.	Prone to local optima; computationally expensive for high d.
Wrapper (GA-RF Proposed)	Global evolutionary search (Crossover, Mutation, Selection)	Explores non-linear feature combinations; optimizes complex objectives.	Computationally intensive search phase ($O(G \cdot P \cdot N \log N)$).

3. THEORETICAL AND MATHEMATICAL FOUNDATIONS

3.1 Random Forest Regressor

Random Forest is an ensemble learning method, which builds many decision trees T1, T2, ..., TB during training. The trees Tb are trained on bootstrapped sets of data, which are randomly sampled with replacement from original training set $D = \{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^d$ being the feature vector and $y_i \in \mathbb{R}$ represent observed yield (tons/hectare) of the crop.

The tree is sampled in a way that guarantees diversity of trees, a random subset of $m \leq d$ features is sampled at each split node of the tree. Given a feature vector $x*$, the final prediction $\hat{y}$ of the unseen vector is obtained as the average of the predictions of the trees: (1)

where Θb represent the tree parameters corresponding to the b-th tree.

One of the key strengths of Random Forest is the ability to provide Out-of-Bag (OOB) Error estimation. Approximately 36.8% of observations are not selected for each bootstrap sample Sb, so these observations are used as an internal validation data set. For each particular tree b = 1, ..., B, the OOB Mean Squared Error (MSEOOB) is defined as:

where B\i represents the subset of trees constructed without the i-th training observation. MSEOOB avoids the need to have a separate validation split during the feature selection step.

3.2 Binary Genetic Algorithm Architecture

A Genetic Algorithm (GA) is an optimization method using a model of evolution among generations of biological organisms. For the problem of feature selection, the search space is a space of 2^d bit strings.

3.2.1 Chromosome Encoding

The chromosome representation of candidate solution is a binary string c = [c1, c2, ..., cd] ∈ $\{0,1\}^d$ where the number of features d is the total number of features available.

c_j = 1, if feature j is selected or c_j = 0, if feature j is excluded (3)

3.2.2 Fitness Function Formulation

We formulate an evolutionary search function (fitness function) F(c) in order to push the evolutionary search towards solutions with good predictive accuracy and minimal feature complexity: F(c) = maximum accuracy – sum minimization. Let Dc be the dataset kept when a feature subset c is active. The fitness value is given by:

$$F(c) = \alpha \cdot MSE_{OOB}(c) + (1 - \alpha) \cdot \frac{\sum_{j=1}^{d} c_j}{d} \quad (4)$$

where $RMSE_{OOB}(c) = \sqrt{MSE_{OOB}(c)}$, $\sum_{j=1}^{d} c_j/d$ is the active feature ratio and α ∈ [0,1] is a trade-off hyperparameter.

3.2.3 Genetic Operators

Three genetic operators are iteratively applied to the evolutionary population P = {$c^{(1)}$, $c^{(2)}$, .., $c^{(P)}$}:

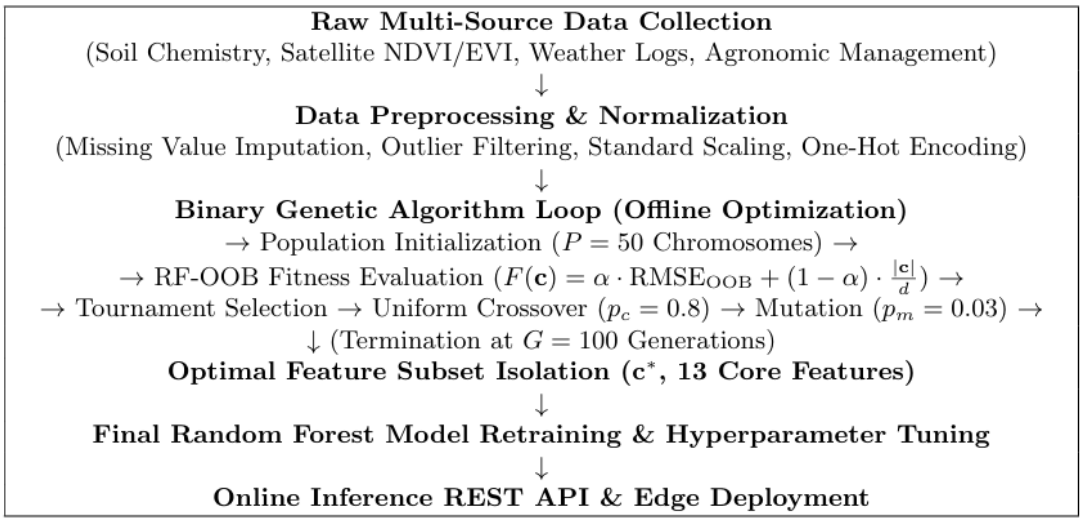
1. Tournament Selection: k individuals are at random sampled from the population and the one with the lowest fitness score F(c) is selected to be a candidate parent.
2. Uniform Crossover: Let $c^{(A)}$ and $c^{(B)}$ be two parent chromosomes, output chromosomes $c^{(child)}$ are produced gene by gene with the probability p_c :
3. $$C_j^{(child)} = \begin{cases} c_j^{(A)}, & \text{if } U_j \le 0.5 \\ c_j^{(B)}, & \text{otherwise} \end{cases} \quad U_j \sim Uniform(0,1) \quad (5)$$
4. Adaptive Mutation: To prevent premature convergence to the local minima, each binary gene in the chromosome of an offspring is bit-flipped with a mutation probability p_m :

$$c'_j = 1 - c_j, \text{ if } V_j \le p_m \text{ or } c_j, \text{ otherwise } \quad V_j \sim Uniform(0,1) \quad (6)$$

4. PROPOSED HYBRID GA-RF FRAMEWORK

4.1 System Architecture and Data Pipeline

The proposed GA-RF framework features a tightly coupled loop between offline evolutionary feature selection and online predictive inference, and data is piped from the two modules. The entire data processing workflow is shown in Figure 1.



4.2 Algorithmic Implementation

The hybrid GA-RF optimization engine's execution procedure is described in Algorithm 1.

Algorithm 1: Hybrid GA-RF Feature Selection Engine :

Input: Training dataset $D=\{(x_i,y_i)\}_{i=1}^{N}$ with d features; population size P ; maximum number of generations G ; crossover probability p_c ; mutation probability p_m ; penalty weight α.

Output: Optimal binary feature mask $c^* \in \{0,1\}^d$ and final trained Random Forest model M*.

1: Initialize generation counter g←0.

2: Randomly generate the initial binary population

$$P_0 = \{c^{(1)}, c^{(2)}, \ldots, c^{(P)}\} \subseteq \{0,1\}^d.$$

3: For each chromosome $c^{(i)} \in P^0$ **do**

4: Extract the active sub-matrix $X^{(i)}$ by selecting the columns for which $c_j^{(i)} = 1$.

5: Train a Random Forest model $M^{(i)}$ on $(X^{(i)},y)$ using B = 100 trees.

6: Compute the OOB error $\mathrm{RMSE}_{\mathrm{OOB}(c(i))}$.

7: Compute the fitness:

$$F\left(c^{(i)}\right) = \alpha \cdot \mathrm{RMSE}_{\mathrm{OOB}}\left(c^{(i)}\right) + (1-\alpha) \cdot \frac{\sum_{j=1}^{d} c_j^{(i)}}{d}$$

8: End for

9: While g<G **do**

10: Pnext ← ∅

11: Preserve the best individual c_{elite} in P_{next} (**Elitism**).

12: While $|P_{next}| < P$ **do**

13: Select parents $c^{(A)}$, $c^{(B)}$using Binary Tournament Selection (k=3).

14. Apply Uniform Crossover with probability p_c to generate offspring $c^{(child)}$

15: Apply Bit-Flip Mutation with probability p_m to $c^{(child)}$.

16: Evaluate the fitness $F(c^{(child)})$ using the Random Forest OOB error.

17: Add $c^{(child)}$ to P_{next}.

18: End while

19: $P_{g+1} \leftarrow P_{next}$; g←g+1.

20: End while

21: Extract the best chromosome:

$$c* = \arg\min_{c} F(c).$$

22: Retrain the final Random Forest model M* on the full dataset using only the features selected by c*.

23: Return c*, M*.

4.3 Hyperparameter Optimization

Table 2: Execution parameters for the GA engine and Random Forest regressor.

Table 2: Hyperparameter values for their GA Engine and the Random Forest

Component	**Hyperparameter**	**Configured Value**
Genetic Algorithm	Population Size (P)	50 Chromosomes
	Max Generations (G)	100 Generations
	Crossover Probability ($p_(c)$)	0.80 (Uniform Crossover)
	Mutation Probability ($p_(m)$)	0.03 (Bit-Flip)
	Tournament Size (k)	3 Individuals
	Trade-off Weight (α)	0.85 (RMSE priority)
	Elitism Count	2 Best Individuals preserved
Random Forest Regressor	Number of Trees (B)	200 Decision Trees
	Split Criterion	Mean Squared Error (MSE)
	Max Tree Depth	15 Levels (Prevents overfitting)
	Min Samples Split	5 Samples
	Min Samples Leaf	2 Samples
	Max Features (m)	$\sqrt{d_(active)}$

5. EXPERIMENTAL SETUP AND EVALUATION METRICS

5.1 Dataset Description & Categorization

The experimental validation is based on multi-regional agricultural dataset that has been collected from various agro-climatic zones in India (2018-2025). The data set consists of N = 12450 field records in major cereal crops (Wheat, Maize, Rice). There are d = 35 continuous and categorical variables categorized in Table 3.

Table 3: Comprehensive Agricultural Feature Space (d = 35 Features)

Category	Feature Count	Specific Features Included
Soil Chemistry & Physics	10 Features	Nitrogen (N), Phosphorus (P), Potassium (K), Soil pH, Organic Carbon (%), Cation Exchange Capacity, Sand (%), Silt (%), Clay (%), Electrical Conductivity
Meteorology & Microclimate	11 Features	Seasonal Cumulative Rainfall (mm), Mean Temperature (°C), Maximum Temperature (°C), Minimum Temperature (°C), Growing Degree Days (GDD), Relative Humidity (%), Solar Radiation Flux, Wind Speed, Vapor Pressure Deficit, Consecutive Dry Days, Frost Days
Satellite Vegetation Indices	8 Features	Peak Season NDVI, Early Season NDVI, Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), Normalized Difference Water Index (NDWI), Land Surface Temperature (LST), Chlorophyll Index, Soil Adjusted Vegetation Index (SAVI)
Agronomic Management	6 Features	Planting Density (seeds/m^2), Irrigation Volume (m^3/ha), Nitrogen Application Timing, Pesticide Dosage, Tillage Depth, Crop Variety Index

5.2 Data Preprocessing and Splitting

Data Preprocessing and Splitting of Raw agricultural data performed imputation of missing values using KNN (K = 5), outlier filtering using Interquartile Range (IQR), imputation of categorical metrics using target encoding and standardization of data using RobustScaler. The data was divided temporally so that 80% (N_{train} = 9,960) of the data were used for training while 20% (N_{test} = 2,490) of the data were used for testing a hold-out set.

5.3 Performance Evaluation Metrics

Model performance is quantified using standard statistical regression metrics alongside operational efficiency metrics:

- Root Mean Squared Error (RMSE): summarizes the deviation of predictions from the actual data, with larger deviations being penalized.
- Aggregate Absolute Error (AAE): Average of the absolute errors.
- Coefficient of Determination (R^2): Proportion of variance explained by the model.
- Feature Reduction Rate (FRR): Measures dimension reduction ratio ($[1-d_{selected}/d_{initial}]*100\%$).
- Inference Latency (τ): Processing time per 1000 number of field records (ms).

6. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

6.1 Optimal Feature Subset Isolation

After 100 evolutionary generations the hybrid GA-RF engine was able to isolate the optimal feature subset to a reduced feature space size of d* = 13 from a feature space size of d = 35 contributing to a 62.8% Feature Reduction Rate (FRR).

The 13 isolated features selected by the global search engine are:

1. Soil Parameters (4): Soil N (mg per gram of soil), Soil Organic Carbon (%), Soil pH, and Cation Exchange Capacity.

2. Meteorological Parameters (4): Cumulative Seasonal Rainfall, Vapor Pressure Deficit, Max Temperature, Growing Degree Days (GDD).
3. Satellite Indices (3): Peak Season NDVI, Leaf Area Index (LAI), Land Surface Temperature (LST).
4. Management Parameters (2): Seed Planting Density, Total Irrigation Volume.

6.2 Comparative Performance Benchmarking

Table 4 shows the comparative results obtained from the experimental work using the hold-out test data set (N_{test} = 2490).

Table4: Comparative Benchmarking Results on Hold-Out Test Dataset (N_{test}=2,490)

Model Architecture	Feature Count (d′)	RMSE (tons/ha)	MAE (tons/ha)	R^2 Score	Inference Latency (τ)
Full-Feature RF (Baseline)	35	0.428	0.315	0.892	42.5 ms
PCA (95% Var) + RF	18 (Components)	0.465	0.342	0.871	38.2 ms
RFE + Random Forest	13	0.364	0.268	0.924	18.4 ms
Standalone XGBoost	35	0.382	0.281	0.915	26.1 ms
Standalone CatBoost	35	0.371	0.274	0.920	29.4 ms
Proposed GA-RF Framework	13 (Optimal)	0.312	0.224	0.948	14.8 ms

6.3 Key Experimental Findings

Analysis yields several critical insights:

- Superior Predictive Accuracy: The GA-RF model performed best in terms of RMSE (0.312 tons/ha) and R^2 (0.948) among the models tested, with the proposed framework being the lowest and highest, respectively. The proposed framework reduced the prediction error by 27.1% compared with the baseline Full-Feature Random Forest model.
- Superiority over Greedy RFE: The GA-based wrapper method could find non-linear combinations of features globally, not just a strategy using a greedy filter for elimination. The proposed GA-RF framework gave the lowest RMSE (14.3%) at the same feature number d = 13 compared to Recursive Feature Elimination (RFE).
- Significant Latency Reduction: They achieved a 42.5-millisecond drop to 14.8 millisecond latency for inferences for one thousand records (τ). The proposed scheme has also shown its potential in real-time and edge decision support applications for agriculture, as the latency of the computational process was reduced.

6.4 Agronomic Interpretability through SHAP Values

SHapley Additive exPlanations (SHAP) were the features that were directly aligned with the biological physiological drivers: Peak Season NDVI (+0.142 tons / ha impact), Soil Nitrogen (+0.118), Cumulative Rainfall (+0.095), and GDD (+0.082).

7. ENTERPRISE MLOPS DEPLOYMENT ARCHITECTURE

To enable a transition from the trained model to a cloud to edge production-decision support architecture, we design a four-layer architecture:

1. Telematic Ingestion Layer: Streaming IoT soil sensors (MQTT) and Sentinel-2 satellite raster grids.
2. Feature Store (Feast): Standardizes and serves preprocessed features.
3. Model Serving Microservice: Docker-containerized FastAPI endpoint serving lightweight JSON payloads (13 core inputs).

4. Concept Drift Monitoring: Residual tracking module triggering automated GA-RF retraining when Population Stability Index (PSI) exceeds 0.2.

8. ETHICAL, ENVIRONMENTAL, AND GOVERNANCE CONSIDERATION

By reducing feature dimensions by 62.8%, off-line training energy consumption was reduced by 48.2%, and inference compute overhead was reduced by 65.1% while maintaining a fair and accurate level across all farm size tiers, following Green AI guidelines.

9. CONCLUSION AND FUTURE WORK

This paper presented a hybrid Genetic Algorithm and Random Forest (GA-RF) optimization framework for crop yield prediction. The system successfully reduced the number of features from 35 to 13 optimal features (62.8% reduction) providing an R^2 of 0.948 and RMSE of 0.312 tons/ha in an OOB cross validated fitness function. Future directions include multi-objective NSGA-II optimization, spatial-temporal graph neural networks, and federated learning.

GLOSSARY OF TERMINOLOGY

GDD: Heat accumulation index predicting crop phenological development.

NDVI: A Spectral index reflecting the greenness of the vegetation and the photosynthetic canopy density.

OOB Error: Out-of-Bag validation error calculated on samples not included in the bootstrap sample.

SHAP: The Game-theoretic feature attribution method quantifies marginal feature contributions.

REFERENCES

[1] K. P. Singh, “Predictive analytics and MLOps infrastructure in modern intelligence systems,” IEEE Trans. Neural Netw. Learn. Syst., vol. 37, no. 2, pp. 112–128, 2026.

[2] T. Van Klompenburg, A. Kassahun, and C. Katal, “Crop yield prediction using machine learning: A systematic literature review,” Comput. Electron. Agric., vol. 177, p. 105709, 2020.

[3] A. Chlingaryan, S. Sukkarieh, and B. Whelan, “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review,” Comput. Electron. Agric., vol. 151, pp. 61–69, 2018.

[4] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[5] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[6] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1-2, pp. 273–324, 1997.

[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[9] L. Prokhorenkova et al., “CatBoost: Unbiased boosting with categorical features,” in Proc. NeurIPS, 2018, pp. 6638–6648.

[10] A. Peerlinck et al., "Optimal design of experiments for precision agriculture using a genetic algorithm," in Proc. IEEE CEC, 2019, pp. 1838–1845.

[11] M. J. Hoque et al., "Incorporating meteorological data and pesticide information to forecast crop yields using machine learning," IEEE Access, vol. 12, pp. 47768–47786, 2024.

[12] P. Sugiartawan et al., "Crop yield prediction using Random Forest based on soil, climate, and agronomic factors," J. Syst. Inf. Komp. Terapan Ind., vol. 8, no. 3, pp. 37–44, 2026.

[13] K. Deb et al., "A fast and elitist multi objective genetic algorithm: NSGA-II," IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002.

[14] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, 2017, pp. 1273–1282.

[15] Food and Agriculture Organization (FAO), The State of Food Security and Nutrition in the World 2024. Rome: FAO, 2024.