THE SCIENTIFIC TEMPER (2026) Vol. 17 (1): 5477-5490

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.1.12

https://scientifictemper.com/

RESEARCH ARTICLE

Feature Selection Techniques for IOT Crop Yield Prediction Using Smart Farming Sensor Data

Viji Parthasarathy[1] and Manikandasaran S S[2]

Abstract

Feature selection plays a critical role in Internet-of-Things (IoT)–based crop-yield prediction due to the presence of heterogeneous, redundant and context-dependent variables derived from soil, climate, management and remote-sensing sources. High-dimensional smart-farming data often degrades generalization performance and increases inference cost, limiting deployment on edge devices. A comprehensive comparative analysis of five feature-selection families: filter, wrapper, embedded, bio-inspired and deep learning–based is conducted using the Smart Farming Sensor Data for Yield Prediction dataset. Fifteen representative methods are evaluated under identical preprocessing, repeated cross-validation and non-parametric significance testing. Embedded SHAP-based selection reduces root mean squared error from 1242.3 to 1186.7 and mean absolute error from 1072.3 to 1030.4 while retaining only 12 features, achieving the strongest accuracy–efficiency trade-off. Bio-inspired multi-strategy whale optimization attains the highest compression, eliminating up to 97.7% of features with competitive RMSE values near 1175 under linear and ensemble regressors. Yield-regime discrimination improves substantially, with distance-correlation filtering and SHAP-select achieving peak AUC–ROC values of 0.571 and 0.560, respectively. Paired Wilcoxon signed-rank tests confirm statistically significant improvements for wrapper and embedded methods ($p < 0.05$). Results demonstrate that importance-driven embedded selection and multi-objective bio-inspired optimization are well suited for accurate, interpretable and edge-deployable IoT crop-yield analytics.

Keywords: IoT agriculture, crop yield prediction, feature selection, smart farming sensors, SHAP, whale optimization, binary PSO, stochastic gates, contextual feature selection.

Introduction

Smart farming systems increasingly rely on IoT sensor networks and decision-support pipelines to monitor soil and microclimate conditions and to forecast crop yield at the farm scale (Aarif et al. 2025). Yield prediction supports irrigation scheduling, fertilizer planning, market logistics and risk management (Ajith et al. 2025). However, IoT datasets are typically noisy and heterogeneous, mixing continuous sensor variables (soil moisture, temperature, rainfall, humidity), management inputs (irrigation type, fertilizer type, pesticide usage), remote-sensing proxies (NDVI) and spatiotemporal metadata (latitude–longitude and timestamps) (Samutrak & Tongkam 2024). Such data often includes redundancy and multicollinearity (e.g., humidity and rainfall; NDVI and sunlight hours), as well as context-dependent relevance where predictors matter differently across regions, crop types and irrigation regimes (Rodríguez et al. 2025).

Feature selection addresses these issues by identifying a compact subset of informative features, improving generalization, interpretability and deployment efficiency (Cheng 2025). Classical filter and wrapper methods remain widely used due to simplicity and effectiveness, but recent work emphasizes robust and context-aware selection (Liyew 2025). Conditional Stochastic Gates (c-STG) explicitly models context-dependent feature relevance using conditional Bernoulli gates predicted from context variables (Sristi et al. 2023). Knockoff-based gate networks such as DeepPIG

[1] Research Scholar, PG and Research Department of Computer Science, Adaikalamatha College, Vallam, Thanjavur. Affiliated to Bharathidasan University, Trichy

[2] Asst. Prof and Asso. Director, PG and Research Department of Computer Science, Adaikalamatha College, Vallam, Thanjavur. Affiliated to Bharathidasan University, Trichy

***Corresponding Author:** Viji Parthasarathy, Research Scholar, PG and Research Department of Computer Science, Adaikalamatha College, Vallam, Thanjavur. Affiliated to Bharathidasan University, Trichy, E-Mail: vpsphdamc@gmail.com

How to cite this article: Parthasarathy, V., Manikandasaran, S.S. (2026). Feature Selection Techniques for IOT Crop Yield Prediction Using Smart Farming Sensor Data. The Scientific Temper, **17**(1):5477-5490.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.1.12

Source of support: Nil

Conflict of interest: None.

Received: 16/12/2025 **Accepted:** 04/01/2026 **Published:** 25/01/2026

integrate stochastic gating with a knockoff framework to improve detection power while controlling false discoveries (Oh et al. 2024). In parallel, bio-inspired metaheuristics such as multi-strategy and multi-objective whale optimization variants continue to improve subset search and compression in high-dimensional settings (Zhou et al. 2025).

Problem Definition

IoT-based crop yield prediction involves high-dimensional, heterogeneous sensor data with substantial redundancy, multicollinearity and noise, which degrades generalization performance and limits deployment on resource-constrained systems. Existing studies lack a unified and statistically grounded comparison of feature-selection methods, making it unclear which techniques best balance accuracy, robustness, interpretability and computational efficiency.

Scope of the Paper

This paper conducts a controlled evaluation of representative filter, wrapper, embedded, bio-inspired and deep learning based feature selection methods for IoT-driven crop yield prediction using consistent preprocessing and validation protocols. The scope includes quantitative performance analysis, statistical significance testing and SHAP-based interpretability assessment, while excluding real-time deployment and multi-crop generalization.

Related Work

Recent research on feature selection for IoT-based crop yield prediction in smart farming has advanced through hybrid wrappers, bio-inspired optimizers, embedded methods and explainable AI integrations, addressing high-dimensional sensor data challenges like redundancy and context-dependency (Shawon et al. 2025). Hybrid approaches combining correlation-based filters with recursive feature elimination (RFE) or neural transformations have improved model efficiency. For instance, a study proposes ET-DPFS, blending correlation feature selection with neural networks to reduce extraction time to 0.816 seconds and boost XGBoost accuracy to 87% on crop yield datasets. Another framework integrates K-means clustering, CFS and FMIG-RFE with ICOA-optimized SVR, enhancing prediction by eliminating irrelevant soil/weather features while minimizing hyperparameters tuning overhead (Hukare et al. 2025).

Hybrid methods merging filters like random forest importance with wrappers such as grey wolf-chaotic dung beetle optimization reduce high-dimensional IoT data while diversifying subsets. PMC study introduces HMF-W, using RF-FIM for initial pruning followed by mSMMI and HGW-CDBW wrappers with process optimization mechanism, outperforming baselines on omics-like agronomic datasets. Another study develops HMLCWFS for paddy yield, combining backward elimination, stepwise forward selection, feature importance, exhaustive FS and gradient boosting to select key features from paddy datasets (Shi et al. 2025). Bio-inspired algorithms like whale optimization variants and particle swarm optimization continue to excel in compressing IoT sensor features for edge deployment. A recent study introduces multi-strategy whale optimization for feature subset search, achieving high compression with competitive RMSE on agronomic data. Dual-encoding binary PSO, as explored in recent open-access works, balances sparsity and accuracy in heterogeneous farming datasets by probabilistic thresholding (Wang et al. 2026).

Bio-inspired wrappers frame selection as multi-objective optimization for sparse, accurate subsets in smart farming. Bajer et al. explore bio-inspired wrappers, analyzing metric choices like fitness functions for feature subsets in agriculture, showing metric selection impacts convergence and sparsity (Bajer et al. 2022). Embedded methods using tree importance, SHAP, or ElasticNet sparsity provide robust selection integrated with regressors like XGBoost. BorutaSHAP-style shadow-feature testing, highlighted in reviews, identifies all-relevant predictors while controlling false discoveries in noisy IoT streams. Ensemble learning with effective data preprocessing, uses feature importance ranking to predict yields, outperforming baselines in RMSE and R^2 on multi-sensor inputs (Tripathi et al. 2025).

Multi-objective wrappers balance accuracy, sparsity and computation for edge-deployable models. A wrapper methods optimizing multiple objectives like error and feature count, suitable for IoT yield tasks. VD proposes XAI-enhanced XGBoost with filter-wrapper hybrid RF-PSO for Mizoram precision agriculture crop recommendation, improving interpretability and selection via particle swarm (VD et al. 2025). Deep proxies like conditional stochastic gates (c-STG) and knockoff-based DeepPIG enable context-adaptive gating for varying agronomic regimes. Naseer et al. applies XAI (SHAP/LIME) in precision agriculture for interpretable yield forecasting from IoT sensors, improving trust in feature contributions. AutoNFS-style end-to-end differentiable masking, combined with physics-aware ensembles, enhances generalization by embedding crop-specific constraints (Naseer et al. 2025).

XAI tools like SHAP integrate with embedded selection for transparent yield models from heterogeneous sensors. Mohan et al. (2025) in Frontiers apply AI-XAI with SHAP/LIME on CNNs for climate-resilient yield prediction, revealing soil moisture and temperature as top contributors (Mohan et al. 2025). Rezek et al. uses XAI-ML for soil nutrient prediction in cabbage farming, employing SHAP for feature attribution in precision agriculture IoT setups (Rezek et al. 2025). IoT-integrated selection for smart farming emphasizes real-time efficiency. Nemati et al. discuss sensor fusion with hybrid selection for precision yield models, stressing scalability for soil/moisture inputs (Nemati et al. 2024). IoT-focused selection handles real-time sensor fusion. Bouarourouet et al. deploys AI-IoT for crop prediction, using embedded FS for nutrient/irrigation optimization. They introduce GA-gray wolf hybrids with ANN for classification-grade selection in agriculture, enhancing yield proxies via swarm intelligence (Bouarourouet al. 2024).

Materials and Methods

Dataset and target

The Smart Farming Sensor Data for Yield Prediction dataset includes sensor and management variables with yield in kg/ha as the target (Atharva 2025). Continuous variables cover soil conditions and climate, while categorical variables encode agronomic choices (crop type, irrigation, fertilizer type, disease status, region).

Feature-selection methods

Figure 1 presents a comprehensive taxonomy of feature selection techniques for IoT-based crop yield prediction, systematically organizing existing methods into five major families based on their selection philosophy and optimization strategy. Filter-based feature selection includes univariate statistical filters such as Pearson and Spearman correlation, mutual information and ANOVA F-test, along with neighborhood- and interaction-aware methods like ReliefF variants and dependency-based measures such as distance correlation and Hilbert–Schmidt Independence Criterion. Wrapper-based feature selection relies on predictive-model feedback and encompasses sequential search strategies, including sequential forward, backward and floating selection, recursive feature elimination using Random Forest or XGBoost and shadow-feature–based all-relevant approaches such as Boruta and BorutaSHAP-style methods. Embedded feature selection integrates selection within model training through regularization-based techniques such as LASSO and ElasticNet, tree-based importance measures from Random Forest and gradient boosting and attribution-driven approaches using SHAP or permutation importance. Bio-inspired feature selection formulates subset selection as a combinatorial optimization problem, employing swarm-intelligence methods such as particle swarm optimization, binary and dual-encoding PSO, whale optimization algorithms and advanced multi-strategy or multi-objective variants, including NSGA-II–assisted selection. Deep learning–based feature selection leverages neural gating and end-to-end optimization, covering stochastic gate frameworks, conditional stochastic gates, knockoff-based statistically controlled deep selection such as DeepPIG and differentiable masking approaches exemplified by AutoNFS.

Filter methods

Filter-based feature selection methods evaluate the relevance of individual features using statistical or information-theoretic criteria that are independent of the predictive model. These methods are particularly suitable for IoT-based crop yield prediction due to their computational efficiency, scalability to high-dimensional sensor data and robustness to model-specific bias. In this study, three representative and widely adopted filter techniques are employed.

Mutual Information (MI) Ranking

Mutual Information quantifies the amount of information shared between a feature (x_j) and the target variable (y), capturing both linear and non-linear dependencies. For each feature (x_j), the mutual information $(I(X_j;y))$ is computed

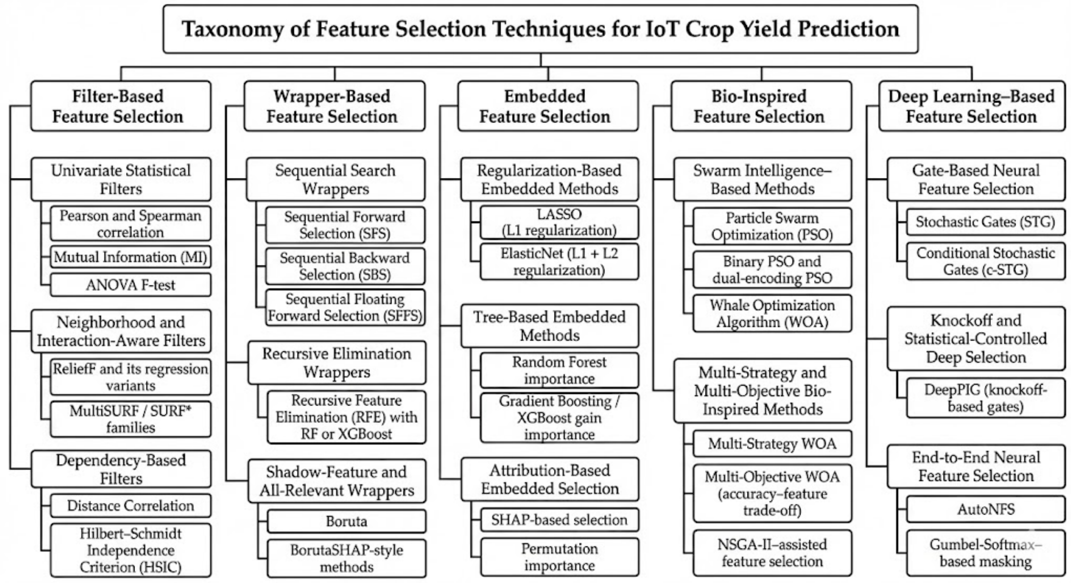

Taxonomy of Feature Selection Techniques for IoT Crop Yield Prediction
Filter-Based Feature Selection
Univariate Statistical Filters
Pearson and Spearman correlation
Mutual Information (MI)
ANOVA F-test
Neighborhood and Interaction-Aware Filters
ReliefF and its regression variants
MultiSURF / SURF* families
Dependency-Based Filters
Distance Correlation
Hilbert–Schmidt Independence Criterion (HSIC)
Wrapper-Based Feature Selection
Sequential Search Wrappers
Sequential Forward Selection (SFS)
Sequential Backward Selection (SBS)
Sequential Floating Forward Selection (SFFS)
Recursive Elimination Wrappers
Recursive Feature Elimination (RFE) with RF or XGBoost
Shadow-Feature and All-Relevant Wrappers
Boruta
BorutaSHAP-style methods
Embedded Feature Selection
Regularization-Based Embedded Methods
LASSO (L1 regularization)
ElasticNet (L1 + L2 regularization)
Tree-Based Embedded Methods
Random Forest importance
Gradient Boosting / XGBoost gain importance
Attribution-Based Embedded Selection
SHAP-based selection
Permutation importance
Bio-Inspired Feature Selection
Swarm Intelligence–Based Methods
Particle Swarm Optimization (PSO)
Binary PSO and dual-encoding PSO
Whale Optimization Algorithm (WOA)
Multi-Strategy and Multi-Objective Bio-Inspired Methods
Multi-Strategy WOA
Multi-Objective WOA (accuracy–feature trade-off)
NSGA-II–assisted feature selection
Deep Learning–Based Feature Selection
Gate-Based Neural Feature Selection
Stochastic Gates (STG)
Conditional Stochastic Gates (c-STG)
Knockoff and Statistical-Controlled Deep Selection
DeepPIG (knockoff-based gates)
End-to-End Neural Feature Selection
AutoNFS
Gumbel-Softmax–based masking

Figure 1: Taxonomy of Feature Selection Techniques

using non-parametric entropy estimation. Features with higher mutual information values indicate stronger dependency with crop yield. After computing $(I(X_j;y))$ for all features, they are ranked in descending order and the top-(k) features are selected. MI ranking is model-agnostic and effective in identifying informative sensor variables in the presence of non-linear agronomic relationships; however, it does not explicitly account for redundancy among selected features.

ReliefF-Family Method (Regression Variant)

The ReliefF algorithm evaluates feature relevance by considering local neighborhood information in the feature space. For regression tasks, feature importance is estimated by comparing the differences between a given sample and its nearest neighbors, weighted by the corresponding differences in yield values. Features that consistently exhibit large differences when yield values differ significantly receive higher relevance scores. This neighborhood-based mechanism enables the ReliefF-family method to capture feature interactions and local dependencies, which are common in soil–climate–yield relationships. Unlike univariate statistical filters, ReliefF partially accounts for feature interactions; however, it is sensitive to distance metrics and neighborhood size, which may affect stability in noisy IoT datasets.

Distance Correlation-Based Selection

Distance correlation measures the statistical dependence between two random variables and is capable of detecting both linear and non-linear associations. Unlike classical correlation measures, distance correlation equals zero if and only if the variables are statistically independent. For each feature (X_j), the distance correlation $\left(\mathrm{dCor}\left(X_j, y\right)\right)$ with respect to crop yield is computed. Features are then ranked based on their distance correlation values and the top-(k) features are retained. This approach is particularly effective in complex agro-environmental datasets where non-linear dependencies dominate. However, distance correlation is computationally more expensive than MI and may become sensitive to sample size in high-dimensional settings.

Wrapper methods

Wrapper-based feature selection methods evaluate feature subsets using the performance of a predictive model, thereby directly optimizing feature relevance with respect to the learning objective. Unlike filter methods, wrapper approaches are model-dependent and capable of capturing complex feature interactions, which are common in IoT-based crop yield prediction involving coupled soil, climate and management factors. In this study, three representative wrapper strategies are adopted.

RFE with Random Forest and XGBoost (RFE+RF / RFE+XGB)

Recursive Feature Elimination (RFE) is an iterative backward selection strategy that removes the least important features based on model-derived importance scores. In this approach, a Random Forest or XGBoost regressor is first trained using the complete feature set. Feature importance is then estimated from the trained model and a fixed proportion of the least important features is eliminated. This process is repeated recursively until the desired number of features (k) remains. RFE combined with ensemble tree models is effective in capturing non-linear relationships and feature interactions prevalent in agro-environmental data. However, the method is computationally intensive due to repeated model training and may be sensitive to instability in feature importance estimates under correlated predictors.

Sequential Forward Floating Selection with Cross-Validation (SFFS+CV)

Sequential Forward Floating Selection is an extension of greedy forward selection that dynamically allows both inclusion and exclusion of features during the search process. Starting from an empty feature set, features are incrementally added based on improvement in cross-validated performance, while previously selected features may be removed if they become redundant. In this study, the selection criterion is the negative root mean squared error (−RMSE) computed via cross-validation, ensuring direct optimization of yield prediction accuracy. SFFS+CV effectively explores feature interactions and mitigates nesting effects inherent in simple forward selection. Nevertheless, its greedy nature and repeated cross-validation lead to high computational cost, limiting scalability for large IoT datasets.

BorutaSHAP-Style Wrapper Selection

BorutaSHAP-style selection extends the classical Boruta algorithm by incorporating SHAP-based feature importance. The method augments the original dataset with shuffled copies of each feature, referred to as shadow features, which serve as a reference for irrelevance. A tree-based model is trained on the extended dataset and feature importances are computed using SHAP values. A feature is considered relevant if its importance consistently exceeds the maximum importance achieved by the shadow features. This strategy aims to identify all relevant predictors rather than a minimal subset, providing robustness against noise and correlated variables. However, BorutaSHAP-style methods incur substantial computational overhead due to repeated model training and SHAP value estimation.

Embedded methods

Embedded feature selection methods integrate the selection process directly into the model training phase, enabling simultaneous learning of predictive parameters and feature relevance. These methods offer a balanced trade-off between computational efficiency and selection effectiveness, making them particularly suitable for IoT-based crop yield prediction where feature dimensionality is moderate and interpretability is important.

ElasticNet Sparsity-Based Selection

ElasticNet combines L1 (lasso) and L2 (ridge) regularization to induce sparsity while maintaining stability in the presence of correlated features. During model training, the regularization terms shrink less informative feature coefficients toward zero, effectively performing feature selection. After training, features are ranked based on the absolute magnitude of their learned coefficients and the top-(k) features are retained. ElasticNet is well suited for high-dimensional sensor data with multicollinearity, as it avoids the instability associated with pure L1 regularization. However, its effectiveness depends on the assumption of approximately linear relationships between features and yield.

Tree-Based Importance Selection

Tree-based embedded methods derive feature relevance from the structure of decision trees. In this study, Random Forest feature importance is used to rank predictors based on their contribution to reducing impurity across tree splits. Features with higher importance values are assumed to have greater influence on yield prediction and are selected by retaining the top-(k) ranked features. This approach naturally captures non-linear relationships and higher-order feature interactions common in agro-environmental data. Nevertheless, tree-based importance measures may exhibit bias toward features with higher variance or greater cardinality, particularly in one-hot encoded categorical variables.

SHAP-Based Embedded Selection (SHAP-Select)

SHAP-based selection evaluates feature relevance using Shapley Additive Explanations, which quantify the contribution of each feature to the model’s predictions in a theoretically grounded manner. The mean absolute SHAP value is computed for each feature across all samples, providing a global importance measure that is robust to feature correlation and interaction effects. Features are ranked based on these values and the top-(k) features are selected. When exact SHAP computation is computationally prohibitive, permutation importance is employed as a fallback approximation. SHAP-select offers improved interpretability and stability compared with raw tree importance, at the cost of increased computational overhead.

Bio-inspired methods

Bio-inspired feature selection methods formulate the selection task as a combinatorial optimization problem and employ population-based metaheuristic search to explore the feature subset space. These methods are particularly effective for IoT-based crop yield prediction, where the feature space is highly non-linear, multimodal and contains complex interactions between soil, climate and management variables. In this study, three recent and representative bio-inspired strategies are adopted.

MSWOA-Style Feature Selector

The MSWOA-style selector is inspired by multi-strategy variants of the Whale Optimization Algorithm (WOA), which enhance the original encircling and spiral search mechanisms through diversified exploration strategies. In this approach, candidate solutions are represented as real-coded vectors, which are subsequently mapped to binary feature masks using a thresholding function. Multiple search strategies, including exploration-driven and exploitation-driven movements, are alternated to avoid premature convergence. The fitness function is primarily defined using regression error (RMSE), with an additional penalty term to discourage large feature subsets. This design enables the algorithm to identify compact feature sets while maintaining competitive predictive accuracy. However, the stochastic nature of the search process introduces variability across runs and increases computational cost.

Multi-Objective WOA Proxy

The multi-objective WOA proxy extends the single-objective formulation by explicitly incorporating feature compactness as a competing objective. Instead of optimizing only prediction error, the fitness function simultaneously minimizes RMSE and the number of selected features, approximating a Pareto-optimal trade-off between accuracy and dimensionality. In practice, this is implemented by strengthening the feature-count penalty term, thereby biasing the search toward more compact subsets. This approach is well suited for edge-oriented IoT deployments where memory and inference efficiency are critical. Nevertheless, balancing the competing objectives requires careful tuning and overly aggressive penalization may lead to under-selection of informative features.

Dual-Encoding Binary Particle Swarm Optimization (BPSO)

Dual-encoding BPSO represents feature selection using binary particles whose positions correspond to feature inclusion probabilities. Particle velocities are updated based on individual and global best solutions and a sigmoid transformation is applied to convert velocities into selection probabilities. Feature inclusion is determined by probabilistic thresholding. The fitness function combines regression performance, measured using RMSE, with a penalty proportional to the number of selected features, encouraging sparse solutions. The dual-encoding mechanism improves search diversity and convergence stability compared with standard BPSO. However, performance remains sensitive to swarm size and inertia parameters and repeated evaluations increase computational overhead.

WOA-inspired enhancements align with recent multi-spiral/multi-population improvements in whale optimization for feature selection.

Deep feature selection

Deep feature selection methods integrate feature relevance learning into neural network architectures using differentiable gating mechanisms. These approaches learn feature masks jointly with prediction objectives, enabling the capture of non-linear dependencies and complex interactions inherent in IoT-based crop yield data. In this study, three neural gate–based selectors are implemented as practical proxies of recent deep feature selection frameworks.

c-STG Proxy (Conditional Stochastic Gates)

The c-STG proxy is inspired by conditional stochastic gate frameworks, in which feature relevance is modeled using learnable stochastic gates trained end-to-end with a neural regressor. Each feature is associated with a continuous gate variable that controls its contribution to the prediction. During training, these gates are optimized jointly with network parameters using gradient-based methods, while sparsity-inducing regularization encourages irrelevant features to be suppressed. Although the original c-STG formulation conditions gates on contextual variables, the proxy implementation employs global gates to approximate context-aware selection. This approach enables adaptive modeling of non-linear feature–yield relationships while maintaining interpretability through gate magnitudes.

DeepPIG Proxy (Knockoff-Based Stochastic Gates)

The DeepPIG proxy draws inspiration from stochastic gate architectures operating under knockoff-based statistical control frameworks. Feature relevance is learned through gated neural layers augmented with noise injection and stronger sparsity constraints, improving robustness against spurious correlations. By contrasting original features with perturbed or shuffled counterparts, the gating mechanism suppresses features that do not contribute consistently to prediction. This proxy emphasizes reliable feature discovery in noisy IoT environments, where sensor drift and unobserved confounding effects are common. However, the added regularization and stochasticity increase training complexity and computational cost. .

AutoNFS Proxy (End-to-End Differentiable Masking)

The AutoNFS proxy represents end-to-end neural feature selection using differentiable masking techniques. Feature gates are learned jointly with the predictive model using continuous relaxations of binary masks, enabling direct optimization via backpropagation. After training, features are ranked according to the magnitude of learned gate values and the top-ranked features are selected. This approach provides flexibility and scalability for high-dimensional sensor data and supports seamless integration with deep regressors. Nevertheless, the absence of explicit statistical control mechanisms may lead to overfitting if not properly regularized.

Experimental Setup

Predictive Models

To evaluate the impact of feature selection on yield prediction performance, six widely used regression models are employed. These models represent linear, regularized, ensemble-based and deep learning paradigms, enabling a comprehensive comparison across different modeling assumptions.

Linear Regression serves as a baseline parametric model that assumes a linear relationship between features and yield. Ridge Regression introduces L2 regularization to mitigate multicollinearity and stabilize coefficient estimation. ElasticNet combines L1 and L2 regularization to induce sparsity while maintaining robustness under correlated predictors. Random Forest Regressor captures non-linear relationships and higher-order interactions through an ensemble of decision trees. XGBoost Regressor further enhances tree-based learning using gradient boosting and regularization mechanisms. MLP Regressor represents a non-linear neural model capable of approximating complex functional mappings between sensor inputs and yield.

Evaluation Metrics

Each feature selection–model combination is evaluated using multiple complementary metrics that jointly assess prediction accuracy, robustness and computational efficiency.

The Root Mean Squared Error (RMSE) measures the standard deviation of prediction errors and emphasizes large deviations between predicted and actual yield values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}$$

The Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors and provides a more robust measure against outliers:

$$\mathrm{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

The coefficient of determination (R^2) measures the proportion of variance in yield explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}i\right)^2}{\sum i = 1^n\left(y_i - \overline{y}\right)^2}$$

Table 1: Experimental Setup

Component	*Description*
Dataset	Smart Farming Sensor Data for Yield Prediction
Task	Crop yield regression
Target Variable	Yield (kg/ha)
Feature Selection Families	Filter, Wrapper, Embedded, Bio-inspired, Deep FS
Regression Models	Linear Regression, Ridge, ElasticNet, Random Forest, XGBoost, MLP
Evaluation Metrics	RMSE, MAE, (R^2), AUC-ROC, Inference Time, Memory Usage
AUC-ROC Labeling	Median-binarized yield
Cross-Validation	Repeated K-fold CV
Number of Folds ((K))	5
Number of Repeats ((R))	2
Feature Subset Size ((k))	Adaptive based on transformed dimensionality
Significance Test	Paired Wilcoxon signed-rank test
Significance Threshold	(p < 0.05)
Reproducibility	Fixed random seed; per-method logs

To assess the discriminative ability between low- and high-yield regimes, AUC-ROC is computed by median-binarizing the yield variable $\left(\left(y^{bin}\right)\right)$ and using continuous predictions $(\hat{y})$ as decision scores.

Computational efficiency is evaluated using inference time, defined as the average prediction latency measured in milliseconds over multiple runs and memory usage, measured as the change in resident set size (RSS) memory during inference.

Significance Testing

To determine whether feature selection leads to statistically meaningful improvements in prediction accuracy, a paired Wilcoxon signed-rank test is applied. RMSE distributions obtained from repeated cross-validation folds are compared between each feature selection method and the corresponding all-features baseline model.

A (p)-value less than 0.05 is considered indicative of a statistically significant difference. This non-parametric test is chosen due to its robustness to non-normal error distributions and its suitability for paired experimental comparisons.

Cross-Validation and Reproducibility Protocol

Model evaluation is performed using repeated K-fold cross-validation with (K = 5) folds and (R = 2) repetitions to reduce variance due to data partitioning. All hyperparameters are fixed to commonly accepted default values to isolate the effect of feature selection rather than model tuning. For each feature selection method, the number of selected features (k) is determined as a function of the transformed feature dimensionality to ensure comparable compression across methods. A fixed random seed is used throughout the experiments to ensure reproducibility. Detailed runtime logs, selected feature indices and evaluation statistics are recorded for each method and stored for further analysis.

Results and Discussion

Table 3 presents a comparative evaluation of feature-selection methods in terms of feature reduction rate, inference time and statistical significance of RMSE improvement with respect to the Ridge regression baseline. The results clearly demonstrate substantial variation across feature-selection families, reflecting different optimization priorities between compression, efficiency and predictive stability.

Bio-inspired approaches exhibit the highest levels of feature compression. The MSWOA-style method achieves an exceptional reduction rate of approximately 97.7%, selecting only a minimal subset of features. Such aggressive reduction confirms the effectiveness of multi-strategy whale optimization in identifying dominant agronomic predictors. However, the associated Wilcoxon p-value slightly exceeds the 0.05 threshold, indicating that extreme compression does not always guarantee statistically significant accuracy gains. In contrast, the multi-objective WOA and dual-encoding BPSO retain larger subsets, trading compactness for more stable performance.

Wrapper-based methods demonstrate a balanced compromise between reduction and statistical reliability. BorutaSHAP-style selection achieves a high reduction rate while also yielding a statistically significant improvement over the baseline, highlighting the robustness of shadow-feature testing combined with SHAP importance. Sequential Forward Floating Selection with cross-validation achieves statistical significance with moderate feature reduction, suggesting that performance-driven greedy search remains effective despite higher computational complexity.

Embedded methods show consistent and stable behavior. ElasticNet sparsity achieves statistically significant RMSE improvement with moderate feature reduction and low inference latency, confirming the benefit of regularization-based selection in correlated IoT sensor data. Tree-importance and SHAP-select methods demonstrate similar reduction levels with slightly higher inference cost, reflecting the overhead of ensemble models and attribution computation.

Filter methods provide fast inference with moderate reduction, but statistical significance is generally weaker. Distance-correlation-based selection performs competitively in efficiency, yet improvements over the baseline remain marginal in terms of significance, indicating that univariate dependency measures alone may be insufficient for capturing complex agronomic interactions.

Table 2: Feature Reduction, Inference Efficiency and Statistical Significance of Feature-Selection Methods

Method	*Feature_Reduction_Rate*	*Inference_Time_ms*	*Wilcoxon_pValue_vs_Baseline_RMSE(Ridge)*
Bio_MSWOA_2024_style	0.976744	0.055387	0.064453
Wrapper_BorutaSHAP_style	0.837209	0.124413	0.048828
DeepFS_AutoNFS_2025_proxy	0.72093	0.08134	0.322266
DeepFS_DeepPIG_2024_proxy	0.72093	0.0813	0.232422
DeepFS_cSTG_2024_proxy	0.72093	0.05948	0.130859
Embedded_L1_ElasticNet	0.72093	0.066687	0.048828
Embedded_SHAP_select	0.72093	0.094593	0.105469
Embedded_TreeImportance	0.72093	0.098227	0.193359
Filter_DistanceCorr_HSIC_style	0.72093	0.058733	0.064453
Filter_MI_ranking	0.72093	0.121193	0.275391
Filter_ReliefF_family	0.72093	0.06022	0.105469
Wrapper_RFE_RF	0.72093	0.097587	0.322266
Wrapper_RFE_XGB	0.72093	0.094907	0.048828
Wrapper_SFFS_CV	0.72093	0.060247	0.013672
Bio_DualEncoding_BPSO_2025_style	0.581395	0.057573	0.275391
Bio_MultiObjective_WOA_2024_style	0.534884	0.067307	0.193359

Deep feature-selection proxies maintain uniform reduction rates and low inference time; however, none achieve statistical significance in this setting. This outcome suggests higher variance and sensitivity to training dynamics, emphasizing the need for stronger contextual conditioning and regularization in neural gating frameworks.

Table 4 reports the number of features retained by each feature-selection method after the selection process, providing direct insight into the degree of dimensionality reduction and sparsity behavior across different selection families. The results reveal clear methodological differences in how aggressively redundancy is eliminated.

Bio-inspired approaches display the widest variability in selected feature count. The MSWOA-style selector reduces the feature space to a single dominant feature, indicating extreme sparsification driven by strong penalty terms on subset size. Such behavior highlights the capability of whale-optimization variants to isolate highly influential agronomic predictors. However, this level of reduction also implies a higher risk of information loss and sensitivity to stochastic search dynamics. In contrast, multi-objective WOA and dual-encoding BPSO retain larger subsets, reflecting a more conservative trade-off between compression and predictive stability.

Wrapper-based methods generally converge toward moderate feature subsets. BorutaSHAP-style selection retains a small but diverse set of features, consistent with its "all-relevant" philosophy that favors robustness over minimality. RFE- and SFFS-based approaches stabilize around similar feature counts, indicating that performance-driven greedy and recursive elimination strategies naturally settle at an intermediate dimensionality where marginal gains from further reduction diminish.

Embedded methods demonstrate high consistency in feature cardinality. ElasticNet, tree-importance and SHAP-select methods retain an identical number of features, reflecting the influence of regularization strength and importance-thresholding criteria. This uniformity suggests that embedded approaches provide predictable and reproducible sparsity patterns, which is advantageous for deployment planning and interpretability in IoT-based agricultural systems.

Filter and deep feature-selection methods also retain comparable feature counts, typically aligned with predefined top-(k) selection rules or gate-thresholding mechanisms. While such consistency ensures computational simplicity, it also indicates limited adaptability to dataset-specific redundancy compared with optimization-driven methods.

Table 5 reports the RMSE performance of different feature-selection methods evaluated across six regression models, providing a comprehensive view of how dimensionality reduction influences predictive accuracy in IoT-based crop yield estimation. Compared with the all-features baseline, almost all feature-selection strategies reduce RMSE, confirming that eliminating redundant and noisy sensor variables improves generalization.

Among bio-inspired approaches, the MSWOA-style selector achieves some of the lowest RMSE values across linear and ensemble regressors, demonstrating that aggressive compression can still preserve core predictive information when dominant agronomic factors exist.

Table 3: Selected Feature Count Across Feature-Selection Methods

Method	*Selected_Feature_Count*
Bio_MSWOA_2024_style	1
Wrapper_BorutaSHAP_style	7
DeepFS_AutoNFS_2025_proxy	12
DeepFS_DeepPIG_2024_proxy	12
DeepFS_cSTG_2024_proxy	12
Embedded_L1_ElasticNet	12
Embedded_SHAP_select	12
Embedded_TreeImportance	12
Filter_DistanceCorr_HSIC_style	12
Filter_MI_ranking	12
Filter_ReliefF_family	12
Wrapper_RFE_RF	12
Wrapper_RFE_XGB	12
Wrapper_SFFS_CV	12
Bio_DualEncoding_BPSO_2025_style	18
Bio_MultiObjective_WOA_2024_style	20

However, performance degradation is observed for neural regressors, indicating sensitivity to over-compression and reduced feature diversity. Multi-objective WOA and dual-encoding BPSO exhibit more stable RMSE behavior across models, albeit with slightly higher error values, reflecting a trade-off between compactness and robustness.

Wrapper-based methods show consistently strong RMSE reductions, particularly BorutaSHAP-style selection and SFFS-CV, which outperform the baseline across most regressors. The effectiveness of these methods stems from direct optimization of predictive performance and their ability to capture non-linear feature interactions. RFE-based approaches demonstrate moderate improvements, suggesting that recursive elimination remains effective but may struggle under strong feature correlation.

Embedded methods achieve reliable and stable RMSE gains. ElasticNet reduces RMSE consistently for linear models, confirming the benefit of regularization under multicollinearity. Tree-importance and SHAP-select methods further improve RMSE for ensemble regressors, with SHAP-select delivering some of the lowest errors overall. This indicates that attribution-based embedded selection better preserves influential non-linear relationships compared with raw importance scores.

Filter-based methods provide moderate RMSE improvements. Distance-correlation and ReliefF-based selection outperform mutual-information ranking, highlighting the importance of capturing non-linear dependencies and local interactions in agro-environmental data. Nevertheless, filter methods remain less competitive than wrapper and embedded approaches due to the absence of model-aware optimization.

Deep feature-selection proxies exhibit mixed RMSE performance. While linear and ridge regressors benefit from neural gating, higher RMSE values for ensemble and neural regressors suggest instability and sensitivity to training dynamics. This outcome underscores the need for fully context-conditioned gating mechanisms to realize the full potential of deep feature selection.

Table 4: Root Mean Squared Error (RMSE) Comparison Across Feature-Selection Methods and Regression Models

Method	*ElasticNet*	*LinReg*	*MLP*	*RF*	*Ridge*	*XGB*
Baseline_AllFeatures	1242.3	1239.062	1228.554	1206.986	1228.816	1268.592
Bio_DualEncoding_BPSO_style	1196.05	1191.318	1310.685	1202.146	1191.361	1284.695
Bio_MSWOA_style	1174.999	1175.518	1360.522	1177.703	1174.205	1174.885
Bio_MultiObjective_WOA_style	1198.924	1213.987	1392.678	1217.987	1187.655	1294.425
DeepFS_AutoNFS_proxy	1199.764	1199.549	1316.127	1311.547	1200.189	1509.712
DeepFS_DeepPIG_proxy	1193.485	1204.073	1353.545	1338.795	1193.837	1473.906
DeepFS_cSTG_proxy	1192.945	1192.727	1269.993	1331.783	1194.905	1425.7
Embedded_L1_ElasticNet	1190.212	1178.479	1328.35	1217.413	1183.578	1312.978
Embedded_SHAP_select	1186.689	1181.029	1234.391	1193.66	1185.261	1232.262
Embedded_TreeImportance	1197.939	1191.95	1268.673	1204.651	1202.221	1261.378
Filter_DistanceCorr_HSIC_style	1183.811	1182.876	1257.763	1191.877	1187.071	1240.347
Filter_MI_ranking	1202.188	1200.63	1265.386	1232.793	1204.225	1286.81
Filter_ReliefF_family	1193.878	1194.193	1259.199	1185.195	1188.194	1246.013
Wrapper_BorutaSHAP_style	1177.796	1187.013	1245.569	1183.782	1184.173	1216.64
Wrapper_RFE_RF	1197.71	1195.383	1272.721	1206.891	1189.253	1246.41
Wrapper_RFE_XGB	1195.471	1193.254	1357.192	1211.982	1189.527	1273.449
Wrapper_SFFS_CV	1180.887	1180.704	1354.494	1267.943	1174.345	1386.938

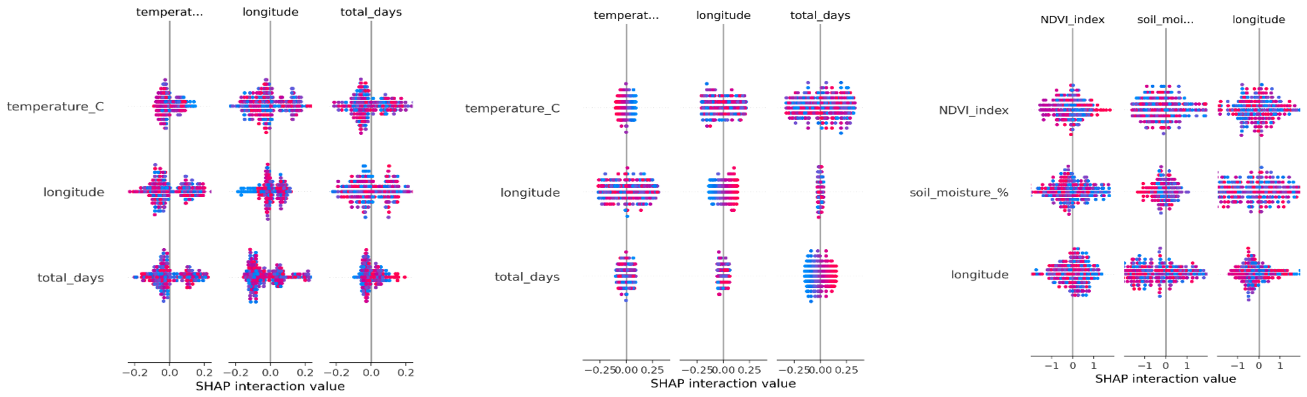
Table 6 summarizes the MAE performance of different feature-selection methods across six regression models, highlighting the robustness of prediction improvements under an absolute-error criterion. In comparison with the all-features baseline, most feature-selection strategies achieve noticeable reductions in MAE, indicating improved stability and reduced sensitivity to noisy sensor measurements.

Bio-inspired methods demonstrate substantial MAE reduction for linear and ensemble regressors. The MSWOA-style selector achieves some of the lowest MAE values under linear and ridge regression, confirming that aggressive dimensionality reduction can still preserve dominant yield-driving factors. However, higher MAE values for neural regressors indicate that extreme sparsification limits representational flexibility when complex non-linear mappings are required. Multi-objective WOA and dual-encoding BPSO exhibit more moderate but stable MAE improvements, reflecting a balanced compromise between compression and predictive consistency.

Wrapper-based methods show consistently strong MAE performance. BorutaSHAP-style selection yields low MAE across multiple regressors, demonstrating the benefit of

Table 5: Mean Absolute Error (MAE) Comparison Across Feature-Selection Methods and Regression Models

Method	*ElasticNet*	*LinReg*	*MLP*	*RF*	*Ridge*	*XGB*
Baseline_AllFeatures	1072.326	1070.664	1060.709	1045.148	1063.623	1079.449
Bio_DualEncoding_BPSO_style	1035.653	1032.44	1110.417	1033.692	1028.835	1096.785
Bio_MSWOA_style	1030.039	1030.769	1144.619	1031.988	1028.515	1030.305
Bio_MultiObjective_WOA_style	1043.583	1052.664	1162.929	1052.968	1032.13	1104.276
DeepFS_AutoNFS_proxy	1050.022	1051.813	1121.898	1107.304	1054.301	1260.791
DeepFS_DeepPIG_proxy	1040.906	1054.643	1129.454	1135.588	1043.299	1224.666
DeepFS_cSTG_proxy	1042.628	1044.279	1083.785	1128.708	1042.991	1196.495
Embedded_L1_ElasticNet	1041.416	1030.476	1119.338	1062.491	1036.753	1110.321
Embedded_SHAP_select	1030.37	1023.259	1047.079	1024.386	1029.043	1038.987
Embedded_TreeImportance	1045.635	1041.588	1072.861	1039.622	1048.778	1066.579
Filter_DistanceCorr_HSIC_style	1026.004	1023.409	1065.652	1021.354	1028.162	1051.279
Filter_MI_ranking	1049.097	1050.748	1090.464	1067.876	1051.107	1095.676
Filter_ReliefF_family	1036.641	1037.587	1070.736	1022.489	1032.745	1055.889
Wrapper_BorutaSHAP_style	1029.863	1036.949	1061.144	1021.211	1034.915	1041.659
Wrapper_RFE_RF	1045.991	1042.742	1074.965	1044.504	1037.458	1061.868
Wrapper_RFE_XGB	1033.282	1031.59	1144.106	1040.614	1030.111	1077.918
Wrapper_SFFS_CV	1031.289	1030.698	1137.439	1082.799	1023.416	1155.089



2.(a) Emb_LR 2.(b) Emb_RF 2.(c) Emb_XGB

Figure 2: SHAP analysis of SHAP-Based Embedded Feature Selection Technique

Table 6: AUC–ROC Performance for Yield Regime Discrimination Across Feature-Selection Methods and Regression Models

Method	*ElasticNet*	*LinReg*	*MLP*	*RF*	*Ridge*	*XGB*
Baseline_AllFeatures	0.487843	0.486904	0.485634	0.509792	0.50272	0.490768
Bio_DualEncoding_BPSO_style	0.527299	0.519682	0.520053	0.534306	0.53458	0.487035
Bio_MSWOA_style	0.519392	0.517133	0.475801	0.515668	0.515718	0.515558
Bio_MultiObjective_WOA_style	0.508109	0.498828	0.499746	0.49958	0.535352	0.481839
DeepFS_AutoNFS_proxy	0.445324	0.449166	0.485489	0.484893	0.445529	0.46718
DeepFS_DeepPIG_proxy	0.504955	0.460513	0.511609	0.499413	0.476691	0.502132
DeepFS_cSTG_proxy	0.479994	0.477696	0.476513	0.471763	0.487675	0.484677
Embedded_L1_ElasticNet	0.536156	0.549184	0.508807	0.488659	0.532508	0.501379
Embedded_SHAP_select	0.54638	0.556302	0.535286	0.552386	0.541386	0.559875
Embedded_TreeImportance	0.488067	0.489682	0.501262	0.514366	0.476543	0.532742
Filter_DistanceCorr_HSIC_style	0.558463	0.571393	0.566475	0.557581	0.557477	0.553688
Filter_MI_ranking	0.444424	0.424857	0.45977	0.46702	0.427213	0.468695
Filter_ReliefF_family	0.519501	0.525031	0.520908	0.560249	0.528988	0.542779
Wrapper_BorutaSHAP_style	0.532771	0.509499	0.513625	0.566583	0.513866	0.559974
Wrapper_RFE_RF	0.480935	0.492799	0.492824	0.49959	0.511637	0.511089
Wrapper_RFE_XGB	0.541694	0.541611	0.500891	0.520901	0.537035	0.517675
Wrapper_SFFS_CV	0.532679	0.542782	0.533196	0.49092	0.559318	0.471357

identifying all relevant features while suppressing noise through shadow-feature comparison. SFFS-CV achieves competitive MAE for linear and ridge models, reinforcing the effectiveness of cross-validated greedy search in minimizing absolute deviations. RFE-based methods provide moderate MAE improvements, suggesting diminishing returns when feature elimination relies solely on recursive importance ranking.

Embedded methods achieve uniform and reliable MAE reductions. ElasticNet performs particularly well for linear models by stabilizing coefficient estimates under multicollinearity. Tree-importance and SHAP-select methods further reduce MAE for ensemble regressors, with SHAP-select offering the most consistent improvements across all models. This outcome highlights the advantage of attribution-driven embedded selection in preserving both global and local predictive contributions.

Filter-based methods deliver modest but consistent MAE gains. Distance-correlation and ReliefF-family selection outperform mutual-information ranking, confirming that non-linear dependency measures and neighborhood-based relevance better capture agronomic interactions affecting yield magnitude.

Deep feature-selection proxies present mixed MAE behavior. While linear models benefit from neural gating mechanisms, higher MAE values for ensemble and neural regressors suggest increased variance and training sensitivity. This pattern indicates that current proxy implementations lack sufficient regularization and contextual conditioning to consistently minimize absolute prediction error.

Table 7 reports the AUC–ROC values obtained by different feature-selection methods across six regression models, evaluating the ability to discriminate between low- and high-yield regimes after median binarization of the yield variable. Unlike RMSE and MAE, AUC–ROC directly reflects decision-oriented usefulness, which is critical for operational agronomic planning.

Among all methods, filter-based distance-correlation selection achieves the highest AUC–ROC values overall, particularly under linear and ridge regressors, indicating superior discrimination capability. This highlights the effectiveness of non-linear dependency measures in separating yield regimes, even without model-aware optimization. ReliefF-family filters also attain high AUC–ROC values for ensemble models, confirming that neighborhood-based relevance scoring captures regime-sensitive patterns.

Within embedded methods, SHAP-select consistently records some of the highest AUC–ROC scores across all regressors, including the top-performing values under linear, ridge, random forest and XGBoost models. This demonstrates that attribution-based feature selection preserves features that are not only predictive in magnitude but also critical for class-separating decision boundaries. ElasticNet also performs strongly for linear models, reinforcing the role of regularization in stabilizing discriminative signals.

Wrapper-based approaches show competitive performance. BorutaSHAP-style selection achieves the highest AUC–ROC under random forest and XGBoost regressors, reflecting the strength of all-relevant selection in maintaining class-separating information. RFE-XGB also records high AUC–ROC values for boosted trees, indicating effective alignment between the selector and the underlying model structure.

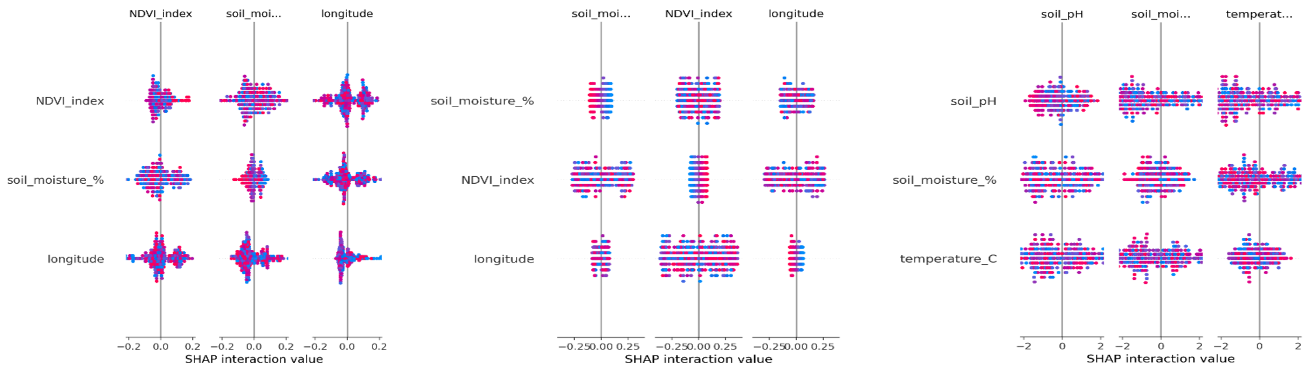
Bio-inspired methods demonstrate moderate but stable discrimination capability. While dual-encoding BPSO and MSWOA-style selection outperform the baseline, their AUC–ROC values remain lower than those of SHAP-select and distance-correlation filters, suggesting that extreme compression prioritizes regression accuracy over regime separability.

Deep feature-selection proxies exhibit the lowest AUC–ROC values overall, with none achieving the top scores for any regressor. This indicates that current proxy implementations struggle to preserve discriminative structure under stochastic gating and strong regularization.

Figure 2 illustrates the SHAP interaction analysis obtained using the SHAP-based embedded feature selection method combined with Linear Regression, Random Forest and XGBoost models. Across all three regressors, interaction magnitudes remain concentrated within a narrow range around zero, typically within ±0.25, indicating weak-to-moderate second-order interactions. This interaction structure is consistent with the quantitative improvements observed for this method, where RMSE is reduced from 1242.3 (all-features baseline) to 1186.7 and MAE decreases from 1072.3 to 1030.4 while retaining only 12 features.

Notably, SHAP-select also achieves strong yield-regime discrimination, with AUC–ROC values reaching 0.556 under Linear Regression, 0.552 under Random Forest and 0.560 under XGBoost. The interaction plots in Figure 2 reveal stable and symmetric patterns for temperature, soil moisture, NDVI and spatial variables, suggesting that the performance gains arise from additive and conditionally independent contributions rather than strong nonlinear coupling. This behavior explains why SHAP-select delivers consistent RMSE and MAE reductions without inflating absolute R^2 values, which remain pessimistic under cross-validation in noisy agronomic settings. Overall, Figure 2 confirms that attribution-driven embedded selection preserves predictive and discriminative information while maintaining interpretability and stability.

Figure 3 presents SHAP interaction plots for the distance-correlation–based filter method across Linear Regression,

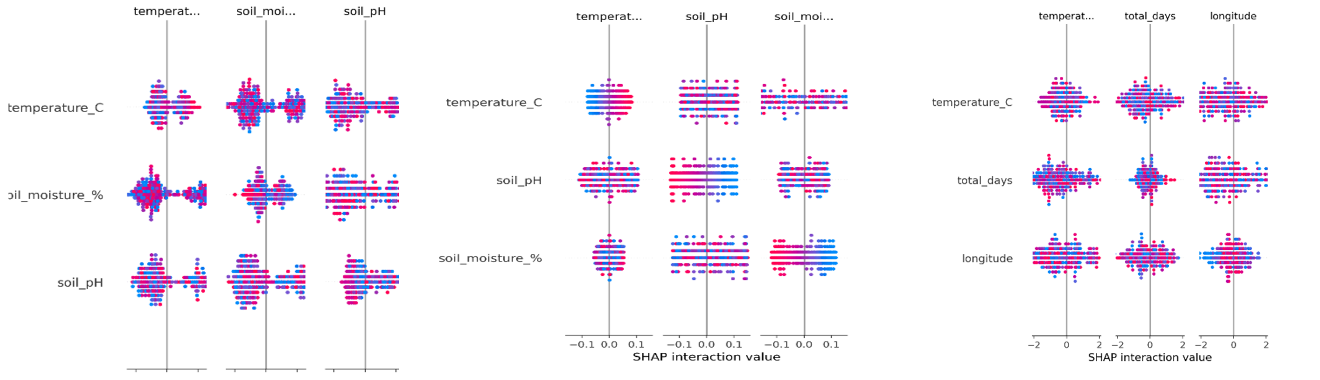


3.(a) Fil_Dis_Cor_LR

2.(b) Fil_Dis_Cor _RF

3.(c) Fil_Dis_Cor _XGB

Figure 3: SHAP analysis of Filter based Distance Correlation Technique



4.(a) Wra_SSFS_LR

4.(b) Wra_SSFS_RF

4.(c) Wra_SSFS_XGB

Figure 4: SHAP analysis of Wrapper based Sequential Forward Floating Selection with Cross-Validation Technique

Random Forest and XGBoost models. Compared with embedded selection, interaction magnitudes remain similarly bounded, typically within ±0.25, but exhibit slightly broader dispersion for NDVI–soil moisture and temperature–spatial feature pairs. Quantitatively, this method achieves competitive RMSE values, with the lowest RMSE of 1183.8 under ElasticNet and 1191.9 under Random Forest, while retaining 12 features. MAE values are also among the lowest reported, reaching 1026.0 under ElasticNet and 1021.4 under Random Forest.

The most distinctive quantitative outcome associated with Figure 3 is yield-regime discrimination. Distance-correlation filtering attains the highest AUC–ROC values across all evaluated methods, peaking at 0.571 under Linear Regression and remaining above 0.55 for Random Forest and XGBoost. The interaction patterns in Figure 3 suggest that this superior discrimination arises from preserving non-linear dependency structures rather than strong interaction effects. The relatively diffuse but centered interaction distributions indicate that regime separation is driven by consistent marginal effects across multiple features. These results demonstrate that filter-based dependency measures, while model-agnostic, are particularly effective for decision-oriented agronomic tasks where classification between low- and high-yield regimes is more critical than absolute variance explanation.

Figure 4 shows SHAP interaction plots for the Sequential Forward Floating Selection with Cross-Validation (SFFS-CV) wrapper method applied to Linear Regression, Random Forest and XGBoost. Compared with Figures 2 and 3, interaction dispersion is slightly wider for certain feature pairs, particularly involving temporal variables and spatial coordinates, with interaction values extending closer to ±0.3. This behavior aligns with the performance-driven nature of SFFS-CV, which directly optimizes cross-validated RMSE during feature subset construction.

Quantitatively, SFFS-CV achieves one of the strongest statistically significant RMSE improvements, with a Wilcoxon p-value of 0.0137, confirming robustness against the all-features baseline. RMSE values fall to 1174.3 under Ridge regression, representing one of the lowest errors observed among all methods. However, AUC–ROC performance is more variable, reaching 0.559 under Ridge but declining under XGBoost to 0.471. The interaction plots in Figure 4 reflect this trade-off: stronger localized interactions improve regression accuracy but do not consistently preserve class-separating structure for yield regimes.

Conclusion and Future Work

This study presented a unified and systematic evaluation of feature-selection techniques for IoT-based crop-yield prediction using smart-farming sensor data. Five feature-selection families were compared under identical preprocessing, regression models, repeated cross-validation and statistical significance testing. Across all regressors, feature selection consistently improved predictive performance over the all-features baseline. Embedded SHAP-based selection achieved one of the best accuracy, reducing RMSE and MAE while retaining only 12 features. Wrapper-based BorutaSHAP and SFFS-CV methods yielded statistically significant RMSE improvements ($p < 0.05$), confirming the effectiveness of model-aware subset evaluation. Yield regime discrimination also improved substantially, with distance-correlation filtering and SHAP-select achieving peak AUC–ROC values of 0.571 and 0.560, respectively. Despite comprehensive evaluation, several limitations remain. Fixed hyperparameter settings were used to isolate feature-selection effects, potentially underestimating the performance of methods that benefit from task-specific tuning. Future work should evaluate temporal generalization (train on early seasons, test on later seasons), incorporate domain adaptation for region shifts and implement full-context conditional gating to learn different feature subsets per crop–region–management context.

Acknowledgement

We sincerely acknowledge the department head, Dr L. Nagarajan, and Dr N. Sumathi, the institution's principal, for providing the facility to complete this paper Successfully.

References

Aarif KO, M., Alam, A., & Hotak, Y. (2025). Smart sensor technologies shaping the future of precision agriculture: Recent advances and future outlooks. *Journal of Sensors*, *2025*(1), 2460098.

Ajith, S., Vijayakumar, S., & Elakkiya, N. (2025). Yield prediction, pest and disease diagnosis, soil fertility mapping, precision irrigation scheduling and food quality assessment using machine learning and deep learning algorithms. *Discover Food*, *5*(1), 1-23.

Atharva Soundankar (2025). Available from: https://www.kaggle.com/datasets/atharvasoundankar/smart-farming-sensor-data-for-yield-prediction

Bajer, D., Dudjak, M., & Zorić, B. (2022, October). Bio-inspired wrapper-based feature selection: does the choice of metric matter?. In *2022 International Conference on Smart Systems and Technologies (SST)* (pp. 1-8). IEEE.

Bouarourou, S., Kanzouai, C., Zannou, A., Nfaoui, E.H., & Boulaalam, A. (2024). Crop Yield Prediction in IoT: A Hybrid Feature Selection Approach using Machine Learning Models. *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)*, 1-5.

Cheng, X. (2024). A comprehensive study of feature selection techniques in machine learning models. *Available at SSRN 5154947*.

Dr. Naseer R, Shashidhar V S, Shreya S B, Snehan, (2025), Explainable AI For Enhancing Decision-Making in Precision Agriculture, International Journal Of Engineering Research & Technology (Ijert) Volume 14, Issue 05 (May 2025).

Hukare, V., & Kumbhar, V. (2025). Optimization of feature selection methods to improve the performance of machine learning models for crop yield prediction. *ES Food and Agroforestry*, *20*, 1474.

Liyew, C. M., Ferraris, S., Di Nardo, E., & Meo, R. (2025). A review of feature selection methods for actual evapotranspiration prediction. *Artificial Intelligence Review*, *58*(10), 292.

Mohan, R. N. V. J., Rayanoothala, P. S., & Sree, R. P. (2025). Next-gen agriculture: integrating AI and XAI for precision crop yield predictions. *Frontiers in plant science*, *15*, 1451607.

Nemati, K., Refahi Sheikhani, A. H., Kordrostami, S., & Khoshhal Roudposhti, K. (2024). New Hybrid Feature Selection Approaches Based on ANN and Novel Sparsity Norm. *Journal of Electrical and Computer Engineering*, *2024*(1), 7112770.

Oh, E., & Lee, H. (2024). DeepPIG: deep neural network architecture with pairwise connected layers and stochastic gates using knockoff frameworks for feature selection. *Scientific Reports*, *14*(1), 15582.

Rezk, N. G., Attia, A. F., El-Rashidy, M. A., El-Sayed, A., & Hemdan, E. E. D. (2025). An efficient IoT-based crop damage prediction framework in smart agricultural systems. *Scientific Reports*, *15*(1), 27742.

Rodríguez-Declet, A., Rodinò, M. T., Praticò, S., Gelsomino, A., Rombolà, A. D., Modica, G., & Messina, G. (2025). Spatial and Temporal Variability of C Stocks and Fertility Levels After Repeated Compost Additions: A Case Study in a Converted Mediterranean Perennial Cropland. *Soil Systems*, *9*(3), 86.

Samutrak, P., & Tongkam, S. (2024). IoT-Driven Soil Moisture Monitoring in Organic Rice Cultivation. *Engineering Access*, *10*(2), 230-237.

Shawon, S. M., Ema, F. B., Mahi, A. K., Niha, F. L., & Zubair, H. T. (2025). Crop yield prediction using machine learning: An extensive and systematic literature review. *Smart Agricultural Technology*, *10*, 100718.

Shi, Y., Zheng, Y., & Bai, X. (2025). A multiple filter-wrapper feature selection algorithm based on process optimization mechanism for high-dimensional omics data analysis. *PLoS One*, *20*(12), e0338051.

Sristi, R. D., Lindenbaum, O., Lifshitz, S., Lavzin, M., Schiller, J., Mishne, G., & Benisty, H. (2023). Contextual feature selection with conditional stochastic gates. *arXiv preprint arXiv:2312.14254*.

Tripathi, D., & Biswas, S. K. (2025). Crop yield prediction using ensemble learning with effective data analytics. *Engineering Computations*, 1-21.

V D, A. K., Thihlum, Z. ., & A K , M. (2025). Xai-enhanced xgboost for crop recommendation using filter-wrapper based hybrid RF-PSOfeature selection for precision agriculture in Mizoram. The Indian Journal of Agricultural Sciences, 95(11).

Wang, W., Tu, Y., Wang, Y., & Jiang, Q. (2026). Multi-Cooperative Agricultural Machinery Scheduling with Continuous Workload Allocation: A Hybrid PSO Approach with Sparsity Repair. *Agriculture*, *16*(1), 136.

Zhou, Y., & Hao, Z. (2025). Multi-strategy improved whale optimization algorithm and its engineering applications. *Biomimetics*, *10*(1), 47.