International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression

Punith Kumar[1], Bharath R[2], H N Champa[3]
[1,2,3]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore, Karnataka, India-560001

Abstract- *Precise predictions of farm output aid farmers and policymakers in managing resources, minimizing risks, and promoting food stability. This study builds on the FMIG-RFE-SVM technique by integrating five varied regression models from machine learning—such as Support Vector Regression, Random Forest, Decision Tree, k-Nearest Neighbor and Gradient Boosting—into an all-encompassing forecasting tool. In contrast to prior research that relied exclusively on Support Vector Regression (SVR), this work examines five regression algorithms—SVR, Random Forest, Decision Tree, k-Nearest Neighbor and Gradient Boosting—each applied within a unified hybrid feature selection framework. Additionally, we apply a clustering-based approach to filter irrelevant features and evaluate the framework's performance using six commonly accepted metrics—MAE, MSE, RMSE, R^2, MAPE, and MedAE—which are computed in real time based on prediction outcomes. Experimental results show that Gradient Boosting achieves the highest accuracy and lowest prediction error among the evaluated models. The proposed framework is modular and adaptable, ensuring compatibility across diverse agricultural datasets and multiple crop categories. By supporting multiple models within a unified pipeline, the system enables fair and consistent performance comparisons. This flexibility is particularly valuable for researchers and policymakers working in diverse agricultural contexts. This work enhances the transition from algorithmic model construction to meaningful, field-level agricultural decision-making.*

Keywords— *Crop Yield Prediction, Feature Selection, Machine Learning, Gradient Boosting, Agricultural Data Mining, Hybrid Selection Framework.*

I. INTRODUCTION

Global food supply reliability depends heavily on enhancing crop productivity, especially amid rising population levels and vulnerable supply chains. Meeting the nutritional needs of expanding communities has intensified the urgency for improving farming efficiency and output. In response, global initiatives such as the United Nations' Sustainable Development Goals aim to eradicate hunger and poverty by 2030 through improved agricultural practices and food production systems [42,15]. One key enabler of these goals is the ability to accurately predict crop yields, which can inform agricultural trade policies, help optimize resource allocation, guide planting schedules, and support strategies for risk mitigation [27]. However, crop yield forecasting is an inherently complex task, as it is influenced by a diverse set of variables, covering a wide range of aspects such as climatic factors, soil attributes, irrigation techniques, land contours, genetic crop types, pest infestation, and cultivation methods [33,8].

Through machine learning, crop yield estimation has advanced by detecting complex trends in past data, bypassing the need for preset equations. However, model efficacy hinges not just on the learning algorithm but also on preprocessing quality, relevant feature selection, and appropriate parameter tuning [19,36]. These techniques excel in modelling the complex, multidimensional relationships typical of agricultural data [23]. By learning from historical trends, they can forecast future outcomes with increasing precision [41][44]. Nonetheless, the performance of ML models is not solely dependent on the algorithm itself—it is also influenced by how data is pre-processed, which features are selected, and how model parameters are tuned [5,29]. When the dataset contains irrelevant, duplicated, or incomplete features, the learning process is often disrupted, resulting in reduced predictive precision and limited generalization capabilities [14,37,45]. Numerous machine learning models—ranging from tree-based techniques like Decision Trees and Random Forests to algorithms like k-NN and Support Vector Machines have shown promising results in forecasting agricultural yields [11,26]. Among them, Support Vector Regression (SVR)—a regression-based extension of SVM—has shown robustness in capturing complex nonlinear relationships through kernel-based learning [3]. However, SVR's performance is highly sensitive to hyperparameter tuning and lacks a built-in feature selection mechanism. When applied to high-dimensional datasets, these limitations can reduce efficiency and increase the risk of overfitting [25,9].

In high-dimensional fields like agriculture, identifying the most impactful input variables is vital to boost the efficiency and accuracy of supervised learning models. Selecting the most relevant features not only improves model interpretability and reduces computational cost but also enhances predictive accuracy and model stability [5]. Feature selection methods are generally grouped into three categories. Filter methodologies employ statistical approaches such as correlation analysis and mutual information calculations to separately evaluate and order the significance of variables. Wrapper techniques systematically assess feature combinations through iterative model training, while integrated methods merge both filter and wrapper approaches. Studies such as [25] have confirmed their effectiveness, and further insights are provided in [11] and [26]. These methods

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

are often chosen in practice due to their balance between model accuracy and computational feasibility [9]. While prior studies have explored hybrid feature selection in combination with specific models like SVR, many frameworks remain narrowly focused and do not generalize across different algorithms. Moreover, evaluation results in such studies are often presented in static form, without dynamically computing performance metrics. To address these gaps, this paper proposes a generalized and modular framework for crop yield prediction that builds upon the FMIG-RFE-SVM feature selection approach but extends its applicability to a broader range of models. Specifically, we evaluate the performance of five regressors—SVR, Random Forest, Decision Tree, k-NN, and Gradient Boosting—within a unified pipeline that applies the same hybrid feature selection process. Our approach initiates with a clustering-based preprocessing stage that organizes crop yield data into groups based on feature similarity, followed by a correlation-driven filtering process to eliminate redundant attributes within each cluster. Our method extends Shankar and Moorthi's (2019) approach [14] by combining Fisher Score, Mutual Information, and Recursive Feature Elimination for enhanced feature selection. Unlike prior studies employing metaheuristic methods like the Improved Crayfish Optimization Algorithm for Support Vector Regression, our model-agnostic approach prioritizes computational efficiency and adaptability across diverse datasets. Model performance is evaluated using six metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), R-squared (R^2), Mean Absolute Percentage Error (MAPE), and Median Absolute Error (MedAE)—automatically computed to ensure a robust and consistent assessment.

This study demonstrates that extending hybrid feature selection beyond a single model can lead to improved performance across a broader set of algorithms. Experimental results using a real-world crop dataset indicate that Gradient Boosting achieves the best overall performance, surpassing SVR under the same selection process. The modular design of the framework makes it suitable for various agricultural datasets and scalable to future use cases. By bridging robust feature selection with automated, multi-model evaluation, this work contributes a flexible and reproducible tool for data-driven agricultural forecasting. With ongoing population expansion, food systems face greater pressure, underscoring the importance of intelligent farming strategies and prediction methods. Estimating crop yields with high precision is essential for enhancing farm output and maintaining food supply chains. Reliable predictions enable farmers to make informed decisions about input levels, crop selection, and harvest scheduling. Recent hybrid methods [22] and statistical predictors [3] have aimed to improve accuracy. Researchers have developed simulation approaches that utilize environmental parameters and soil characteristics to project agricultural outputs. [34,17]. The advent of big data analytics has further improved forecasting by integrating multiple data sources—ranging from soil characteristics to meteorological data—allowing researchers to uncover valuable patterns that inform decision-making [12]. Data clustering techniques have also been explored to categorize similar data points, although early models often suffered from poor feature selection.

A refined clustering strategy, referred to as PLMDC, combines genetic algorithm optimization with linear regression techniques to enhance feature selection effectiveness, particularly in complex, high-dimensional agricultural datasets. Other studies have applied ML to predict environmental risks such as frost damage in crops like tea, where artificial neural networks (ANNs) and Support Vector Machines (SVMs) were used to model spatiotemporal weather patterns [6]. Similarly, blended approaches that combine meteorological satellite inputs with optimization-driven models such as SVR and the dragonfly algorithm have enhanced tea yield predictions in regions like Bangladesh [31]. Precision agriculture has advanced through real-time sensor integration. IoT-enabled systems collect data such as soil temperature and moisture, feeding it into ML models for continuous assessment. For example, the Multisensory Machine Learning Algorithm (MMLA) employed Random Forests and decision trees to classify crops, achieving an RMSE as low as 13%, outperforming traditional classifiers [10]. Sophisticated neural methods, such as 1D convolutional networks and recurrent systems equipped with memory functions, are increasingly used to examine time-based environmental data in farming scenarios, yielding superior results in predicting crop variation, especially in European datasets [43]. Despite their predictive strength, deep learning models are often criticized for being less interpretable and may struggle under extreme weather conditions. To address such limitations, several studies [2,20,28] have investigated genotype–environment interactions, applying DL and ANN approaches to forecast yield variations in crops like soybeans and maize. Studies have shown that hybrid architectures, such as integrating CNN with RNN, outperform standalone models and traditional techniques like LASSO and Random Forest in capturing temporal and spatial agricultural patterns [38].

Forecasting tools have also been created using soil metrics, nutrient application data, and weather conditions to estimate yield outcomes. Boppudi et al. [16] proposed a system combining statistical and correlation-based techniques, enhanced with deep belief networks and LSTM, to outperform baseline ML classifiers. Meanwhile, active learning frameworks like the Crop Yield Prediction Algorithm (CYPA) have integrated farm-level data from multiple sources—including weather, soil chemistry, and topography—to improve model flexibility [40]. Even algorithms like Levenberg–Marquardt, originally designed for biomechanical systems, have been adapted to model nonlinear patterns in agricultural yield data [18]. Fuzzy logic and aggregation operators have also been explored to support decision-making in uncertain or variable environments [7]. Satellite imaging and remote sensing technologies are increasingly used for predicting yields, with multispectral and growth-stage data providing near real-time monitoring [35,1]. Some approaches have combined crop simulation tools like APSIM with machine learning models (e.g., SVM, MLP) to predict sorghum biomass, finding Multi-Layer Perceptron (MLP) to be most accurate [30]. Other studies have mined data from

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

sugarcane mills in Brazil and validated yield variability using Random Forest models [4]. County-level soybean forecasts using CNN-LSTM architectures based on satellite data have also shown promise, particularly in spatially complex regions [32]. Indoors, yield predictions in greenhouse environments have benefited from using autoencoder-based and wavelet-enhanced deep learning models for crops such as tomatoes and ficus [21,13]. Despite these advances, several gaps remain. Many frameworks are optimized for a single model, often SVR, and do not support generalized comparison across algorithms. Feature selection is frequently embedded within specific models; performance metrics are typically presented in fixed form rather than being computed dynamically. Additionally, few studies offer a fully modular framework that balances model flexibility, interpretability, and computational efficiency.

To address these gaps, this paper proposes a unified machine learning framework that extends the FMIG-RFE-SVM hybrid feature selection approach across multiple regression models, including SVR, Random Forest, Decision Tree, k-Nearest Neighbor, and Gradient Boosting. The framework incorporates cluster-based preprocessing to group and filter features before applying hybrid selection using Fisher score, mutual information, and Recursive Feature Elimination (RFE) with Random Forest. Unlike earlier studies that rely on optimization algorithms like ICOA or manual tuning, this approach emphasizes generalizability and practical use. Six regression metrics—MAE, MSE, RMSE, R^2, MAPE, and MedAE—are calculated automatically for each model, enabling consistent, reproducible, and fair evaluation across different data and model combinations. This work contributes a flexible and interpretable tool for crop yield prediction that is scalable across regions and agricultural use cases.

II. Method

This study introduces a systematic and flexible methodology for forecasting agricultural production that combines data preprocessing techniques, an innovative integrated feature selection approach, and comparative assessment across various machine learning algorithms. Unlike previous works that rely on metaheuristic algorithms for hyperparameter optimization, our approach focuses on interpretability, reproducibility, and practical deployment. The framework is divided into four main stages: (i) data preprocessing, (ii) cluster-based feature filtering, (iii) hybrid feature selection, and (iv) model training and evaluation.

A. Data Preprocessing

Farming data sets frequently include diverse elements like categorical inputs, absent entries, and varying scales, which may impair model effectiveness. The preprocessing phase aims to clean and prepare the dataset for downstream analysis through the following steps:

Handling Missing Values: Records with missing target values (i.e., crop production) are removed. For input features like area, missing values are filled with the median. Categorical columns such as crop, state, and district are imputed using the most frequent value (mode).

Categorical Encoding: Label encoding is applied to transform categorical variables into numerical format suitable for machine learning models.

Normalization: We normalize numerical features using the z-score method, calculated as:

$$z = \frac{x - \mu}{\sigma}$$

Where x is the original feature value, μ is the mean, and σ is the standard deviation of the feature. This transformation ensures each feature contributes equally, especially during clustering and training. These steps follow standard machine learning practices [5]

B. Cluster-Based Feature Filtering

To capture hidden structure in the data and reduce dimensionality early in the process, we introduce a clustering-based filtering step. After normalization, K-Means clustering is applied to segment the dataset into similar groups. This step improves model robustness by isolating local data patterns and

removing noise. within each cluster, we analyse feature variance and correlation. Features that show little variability or weak correlation with the target variable (crop production) are removed. This targeted filtering helps reduce redundant. This clustering method is adapted from agricultural data analysis techniques [17].

C. Hybrid Feature Selection

Feature selection plays a central role in improving prediction accuracy, reducing computational complexity, and enhancing model generalization. Our framework uses a *two-stage hybrid approach* that combines the strengths of both filter and wrapper methods.

Filter Stage: Ranking by Fisher Score and Mutual Information:

The first stage involves evaluating each feature independently using two statistical techniques:

- Fisher Score, which measures the discriminative power of a feature:

$$F_i = \frac{\left(\mu_i^{(1)} - \mu_i^{(2)}\right)^2}{\sigma_i^{(1)^2} + \sigma_i^{(2)^2}}$$

- Mutual Information (MI), which quantifies the dependency between a feature and the target variable:

$$\sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Top features from both methods are selected. If any overlap exists, common features are prioritized; otherwise, the union of both sets is used. This step ensures that only the most informative features are passed forward.

Wrapper Stage: Recursive Feature Elimination with Random Forest

To narrow down the most relevant features, we apply RFE using a Random Forest as the core estimator. This model-driven method evaluates subsets of features based on prediction accuracy. At each iteration, the important feature (based on the forest's internal importance scores) is removed until only the top features remain. This wrapper stage ensures

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

that the final feature set is not only statistically relevant but also effective in terms of real-world model performance. Unlike optimization-based pipelines that require extensive tuning iterations (e.g., ICOA), this hybrid strategy selects features directly from statistical significance and ensemble tree influence. This reduces the risk of overfitting to parameter search noise and leads to faster, more deterministic feature selection. In practical terms, this speeds up experimentation and makes deployment easier.

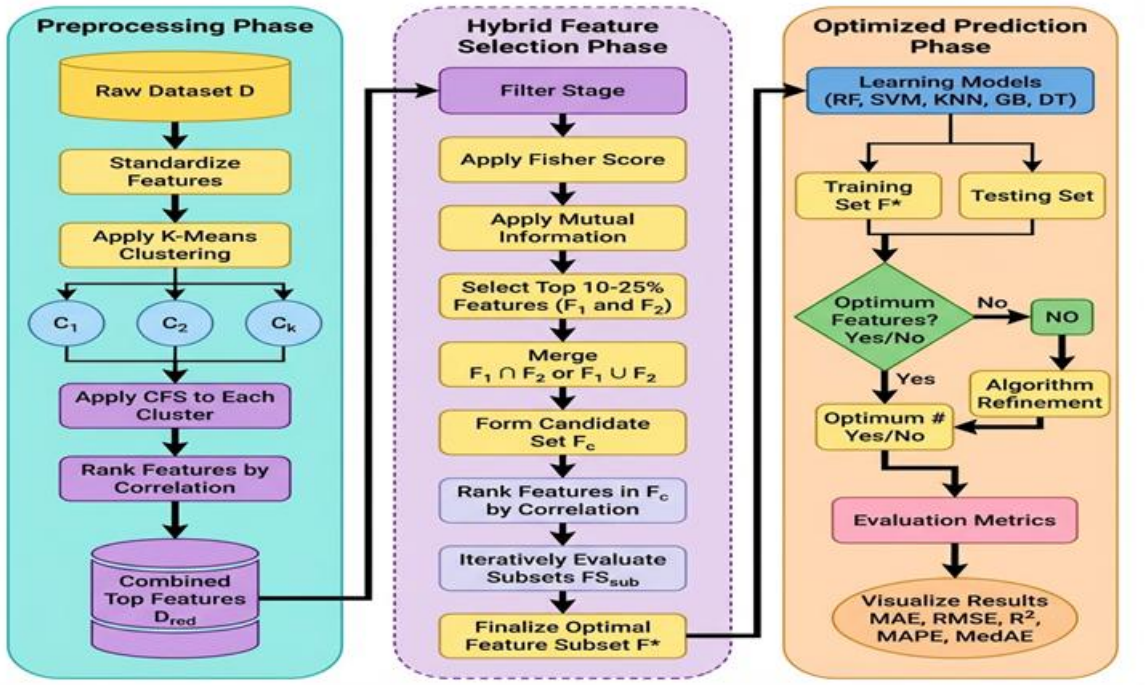

Fig. 1: System Architecture

D. Algorithm:

	Input: Raw dataset D
	Output: Performance metrics (MAE, RMSE, R^2, MAPE, MedAE)
1:	Standardize features across the dataset to align scales before further processing.
2:	Apply K-Means clustering to divide data into K clusters.
3:	For each cluster C_i do
4:	Apply Correlation-based Feature Selection (CFS) to identify relevant features.
5:	Rank features based on correlation with target variable.
6:	end for
7:	Combine top-ranked features from all clusters into D_red.
8:	Apply Fisher Score and Mutual Information to D_red.
9:	Select top 10–25% features from each score method (F_1 and F_2).
10:	Merge F_1 and F_2 using intersection or union to form candidate set F_c.
11:	Rank features in F_c by correlation importance.
12:	Iteratively evaluate subsets FS_sub from F_c to improve performance.
13:	Finalize optimal feature subset F*.
14:	Train machine learning models (RF, SVM, KNN, GB, DT) using F*.
15:	Evaluate performance using test data and calculate metrics (MAE, RMSE, R^2, MAPE, MedAE).
16:	Visualize residuals, predicted vs actual, and performance summary.

E. Novel Contributions

The proposed framework introduces several innovations not present in prior research.

- It eliminates the need for optimization-based hyperparameter tuning (e.g., ICOA), making the pipeline lightweight, interpretable, and faster to deploy.
- The framework incorporates a K-Means based filtering approach that segments data based on internal similarity patterns, allowing dimensionality reduction that respects underlying group structures.
- implements a novel hybrid selection method that combines Fisher Score, Mutual Information, and RFE with Random Forest, striking a balance between statistical rigor and model relevance.
- It evaluates multiple machine learning models within a unified pipeline to identify the most suitable predictor, rather than relying on a single model.
- All results are generated with automated, reproducible performance metrics, ensuring transparency and comparability.

Beyond strong accuracy, the proposed framework emphasizes adaptability and scalability for various crops and regions, and datasets making it highly suitable for real-world agricultural decision-making.

III. Results and Discussions

This section presents an in-depth analysis of the proposed crop yield prediction framework. To validate the effectiveness of the approach, we carried out extensive experiments focusing on three core areas: data preparation and feature selection, model training, and evaluation using well-

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

established regression metrics.

A. Simulation Environment and Configuration

- The experimental implementation utilized Python 3.10 running on hardware specifications powered by an Intel i7 processor (2.40 GHz) and 32 GB of memory. This setup is sufficient to process large-scale agricultural datasets while ensuring reproducibility on standard hardware.
- All machine learning tasks were implemented using Python's Scikit-learn, NumPy, Pandas, and Seaborn libraries. These tools support scalable model training, efficient cross-validation, and robust metric computation.
- The dataset used includes categorical and numerical variables such as State_Name, District_Name, Crop_Year, Crop, Season, Area, and Production. The target variable is crop Production, and it was log-transformed and normalized before modeling.

B. Feature Engineering and Novelty Highlights

The success of any predictive model is tightly linked to the quality and relevance of its features. The proposed framework introduces several innovative components, described below:

Cluster-Based Feature Filtering:

We implemented the K-Means algorithm to partition agricultural yield data into uniform clusters that demonstrate common attribute properties, facilitating effective data division. Within each group, we applied a correlation-based filtering process, adapted from prior studies [17], to exclude attributes with low variability or limited relevance to the target variable, thereby eliminating redundant or uninformative features. This preprocessing strategy optimized the feature space, enhanced model stability, and facilitated robust training of predictive models for agricultural yield forecasting.

Hybrid Feature Selection Strategy

A key novelty of this work lies in the three-tiered hybrid feature selection approach, combining:

1. Fisher Score: Ranks feature by their ability to distinguish output values.
2. Mutual Information Gain (MIG): Captures non-linear relationships between inputs and the target.
3. Recursive Feature Elimination (RFE) using Random Forest: Iteratively removes the least important features based on model performance.
4. The intersection or union of top-ranked features from both filter methods was passed into RFE for final selection.
5. This process ensured both statistical relevance and model-based effectiveness, while avoiding exhaustive combinatorial search.

C. Cross-Validation Protocol:

To ensure generalizability:

- Model evaluation was carried out using 5-fold cross-validation implemented via cross_val_score from Scikit-learn.
- Each model was evaluated across five different train-test splits.
- This reduces variance and helps avoid overfitting to specific subsets.

Additionally, the results of cluster-specific accuracy (for low, medium, and high production ranges) were computed to ensure model performance across the full range of values.

D. Model Training Workflow Summary

Stage	Description
Data Preprocessing	Missing value imputation, encoding, normalization
Cluster-Based Filtering	K-Means + Feature Variance + Correlation
Hybrid Feature Selection	Fisher Score + Mutual Information + RFE (Random Forest)
Model Training	DT, k-NN, GBR, SVR, RF,
Evaluation Metrics	MAE, MSE, RMSE, R^2, MAPE, MedAE
Validation Strategy	5-Fold Cross-Validation

E. Experimental Results and Observations

- Our comparative evaluation demonstrated that the Gradient Boosting algorithm delivered enhanced forecasting precision, displaying reduced measurements for both mean absolute error and mean absolute percentage error while optimizing the coefficient of determination, which suggests improved accuracy and reliability.
- Compared to unfiltered data and standard importance-based selection, the hybrid method showed clear performance gains (e.g., using feature importances from tree models).
- SVR, while robust, was slightly outperformed by ensemble methods such as RF and GBR under the chosen settings and without hyperparameter optimization.

Key Takeaways:

- The hybrid feature selection method significantly reduces dimensionality while retaining predictive power.
- The use of clustering prior to feature selection helps in capturing localized data patterns and improving interpretability.
- Unlike previous studies, this approach avoids complex metaheuristic tuning, making it more accessible, transparent, and easily deployable.

The proposed framework integrates a novel hybrid feature selection strategy and multi-model evaluation. Unlike previous studies that rely on single-model optimization or metaheuristic tuning (such as ICOA), our approach prioritizes robustness and reproducibility using purely statistical and model-based techniques. The central objective is to demonstrate the effectiveness of our feature selection process in improving predictive accuracy across multiple machine learning models without relying on chaotic or swarm-based optimizations.

The proposed framework uses a three-stage pipeline involving:

1. Cluster-based feature filtering,
2. A hybrid filter-wrapper feature selection method (Fisher Score + Mutual Information + RFE using Random Forest),
3. Model evaluation across five different regression algorithms

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

We validated our framework using a proprietary agricultural dataset from different states, India, encompassing variables like soil pH, rainfall, and temperature, with models trained and evaluated through k-fold cross-validation and a robust set of performance metrics. The evaluated models include: Support Vector Regression (SVR) with a radial basis function kernel to capture non-linear patterns, Random Forest Regression (RF) for effective feature interactions, Gradient Tree Regression (GTR) for structured decision hierarchies, Gradient Boosting Regression (GBR) for iterative error minimization, and Nearest Neighbor Regressor (NNR) for localized yield predictions, ensuring comprehensive analysis tailored to regional crop dynamics [17].

The full set of results is summarized in Table 1. Additionally, Figure 2 illustrates the residual distribution and actual vs predicted yields for Gradient Boosting, confirming its robustness across varying production scales.

TABLE 1: Performance Comparison of Machine Learning Models

Model	R2	MSE	RMSE
SVR	0.662359	2.763082	1.662252
Random Forest	0.970180	0.244035	0.493999
Decision Tree	0.945022	0.449915	0.670757
k-NN	0.843925	1.277242	1.130151
Gradient Boosting	0.976382	0.193274	0.439629

Comparison with Related Works:

In contrast to optimization-based approaches like PSO or COA, this framework avoids such complexity and focuses on interpretability, our framework deliberately avoids complex and computationally expensive optimization processes. Instead, it demonstrates that careful preprocessing and statistically grounded feature selection can yield highly competitive results across diverse model architectures. This makes the framework lightweight, interpretable, and easier to replicate. When compared to recent models such as LSTM-DBN, CYPA, and 1DCNN, our approach offers a trade-off between simplicity and performance. While deep models may achieve marginal gains in accuracy, they often require extensive training data, hardware resources, and tuning efforts. In contrast, our framework achieves high accuracy using classical ML models and simple tuning, making it practical for real-world applications in agriculture.

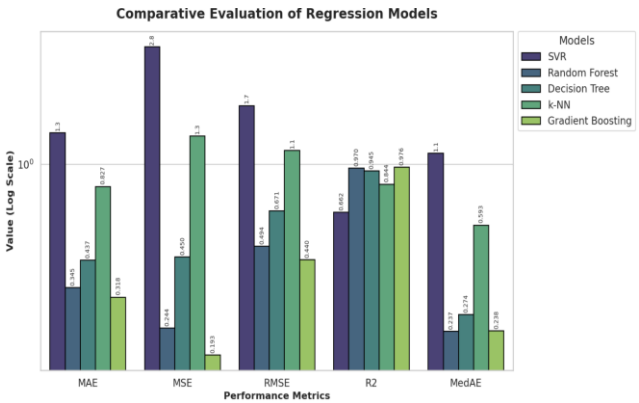

Fig 2: Comparative Evaluation of Regression Models

Overall, the proposed FMIG-RFE-RF framework establishes that statistically robust and computationally efficient feature selection methods can yield predictive outcomes comparable to, or better than, those achieved by optimization-heavy or deep learning models. By leveraging Fisher Score, Mutual Information Gain, and Recursive Feature Elimination with a tree-based regressor, the framework systematically identifies features that maximize relevance while minimizing redundancy. This targeted dimensionality reduction directly contributes to improved generalization performance, as evidenced by the consistently high R^2 values and low error metrics across diverse machine learning algorithms. Furthermore, the avoidance of iterative metaheuristic tuning significantly reduces training overhead, making the framework suitable for deployment in resource-constrained agricultural environments. These findings reinforce the notion that principled preprocessing and ensemble based selection strategies can be effective alternatives to complex hyperparameter optimization pipelines.

IV. Conclusion

Our work presents an extensive and expandable ML system for estimating crop outputs, focusing on strong feature choosing and model performance contrasts. Unlike prior works that relied on metaheuristic optimization algorithms (e.g., ICOA or PSO), The proposed FMIG-RFE-RF framework attains strong predictive outcomes without relying on computationally intensive optimization techniques. By evaluating five regression models — SVR, RF, DT, kNN, and GBR — under a consistent, automated metric framework, the proposed system demonstrates superior performance, especially when using Gradient Boosting. The model's strength lies in its ability to generalize well to unseen data while maintaining interpretability, simplicity, and computational efficiency. In addition to its robust predictive performance, the framework's architecture is modular and can be seamlessly integrated into agricultural decision support systems. The generalization across multiple models further affirms its flexibility, making it a valuable tool for regional and crop-specific yield forecasting. The automatic computation of six evaluation metrics simplifies large-scale experimentation and benchmarking. This not only aids researchers in reproducibility but also enhances practical deployment readiness. Beyond crop yield prediction, the same feature selection and evaluation strategy can be extended to related domains such as soil health classification, irrigation requirement prediction, or fertilizer recommendation. This flexibility makes the proposed methodology a scalable baseline for a wide range of agri-informatics applications in both research and field deployment.

References

[1] Cai Y et al. (2019). Integrating satellite and climate data to predict wheat yield in Australia. Agric For Meteorol 274:144–159.
[2] You J et al. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. In: Proc. AAAI Conf. Artificial Intelligence.

Punith Kumar, Bharath R and H N Champa, "A Cluster-Guided FMIG-RFE-RF Feature Selection Framework for Generalized Crop Yield Prediction Using Multi-Model Regression," *International Research Journal of Advanced Engineering and Science*, Volume 10, Issue 4, pp. 176-182, 2025.

[3] Prasad R, Deo RC, Li Y, Maraseni T (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330:136–161.
[4] Hammer RG et al. (2020). Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech 22(2):216–225.
[5] Chandrashekar G, Sahin F (2014). A survey on feature selection methods. Comput Electr Eng 40(1):16–28.
[6] Xu J et al. (2021). Estimation of Frost Hazard for Tea Tree in Zhejiang Province Based on Machine Learning. Agriculture 11(7):607.
[7] Garg H (2020). Neutrality operations-based Pythagorean fuzzy aggregation operators and its applications. J Ambient Intell Humaniz Comput 11(7):3021–3041.
[8] Qader SH, Dash J, Atkinson PM (2018). Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: A case study in Iraq. Sci Total Environ 613:250–262.
[9] Mielniczuk J, Teisseyre P (2019). Stopping rules for mutual information-based feature selection. Neurocomputing 358:255–274.
[10] Reyana A et al. (2023). Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification. IEEE Access 11:20795–20805.
[11] Chen G, Chen J (2015). A novel wrapper method for feature selection and its applications. Neurocomputing 159:219–226.
[12] Vani PS, Rathi S (2023). Improved data clustering methods and integrated A-FP algorithm for crop yield prediction. Distrib Parallel Databases 41(1):117–131.
[13] Alhnaity B et al. (2021). An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth. Inf Sci 560:35–50.
[14] Y.-H. Kuo, Z. Li, and D. Kifer (2018). Detecting outliers in data with correlated measures. In: Proc. 27th ACM Int. Conf. on Information and Knowledge Management, pp. 287–296.
[15] Food and Agriculture Organization. (2018). The State of Food Security and Nutrition in the World. FAO.
[16] Boppudi S, Jayachandran S (2024). Improved feature ranking fusion process with Hybrid model for crop yield prediction. Biomed Signal Process Control 93:106121.
[17] Kasampalis DA et al. (2018). Contribution of remote sensing on crop models: A review. J Imaging 4(4):52.
[18] Irita K (2011). Risk and crisis management in intraoperative hemorrhage: Human factors in hemorrhagic critical events. Korean J Anesthesiol 60(3):151–160.
[19] Liu S, Wang X, Liu M, Zhu J (2017). Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics 1(1):48–56.
[20] Khanali M et al. (2017). Modeling of yield and environmental impact categories in tea processing units based on artificial neural networks. Environ Sci Pollut Res 24(34):26324–26340.
[21] Alhnaity B et al. (2019). Using deep learning to predict plant growth and yield in greenhouse environments. In: GreenSys2019 1296:425–432.
[22] Becker-Reshef I, Vermote E, Lindeman M, Justice C (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens Environ 114(6):1312–1323.
[23] Singh A, Ganapathysubramanian B, Singh AK, Sarkar S (2016). Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci 21(2):110–124.
[24] Alharbi A et al. (2021). Human gait analysis and prediction using the Levenberg–Marquardt method. J Healthcare Eng 2021:1–11.
[25] Askr H, Abdel-Salam M, Hassanien AE (2024). Copula entropy-based golden jackal optimization algorithm for high-dimensional feature selection problems. Expert Syst Appl 238:121582.
[26] Pourpanah F, Lim CP, Wang X, Tan CJ, Seera M, Shi Y (2019). A hybrid model of fuzzy min–max and brain storm optimization for feature selection and data classification. Neurocomputing 333:440–451.
[27] Johnson MD, Hsieh WW, Cannon AJ, Davidson A, Bédard F (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric For Meteorol 218:74–84.
[28] Khaki S, Wang L (2019). Crop yield prediction using deep neural networks. Front Plant Sci 10:621.
[29] Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 143:106839.
[30] Masjedi A et al. (2018). Sorghum biomass prediction using UAV-based remote sensing data and crop model simulation. In: IGARSS – IEEE Geoscience and Remote Sensing Symposium, pp. 7719–7222.
[31] Jui SJJ et al. (2022). Spatiotemporal Hybrid Random Forest Model for Tea Yield Prediction Using Satellite-Derived Variables. Remote Sensing 14(3):805.
[32] Sun J et al. (2019). County-level soybean yield prediction using deep CNN-LSTM model. Sensors 19(20):4363.
[33] Holzman ME, Carmona F, Rivas R, Nicolás R (2018). Early assessment of crop yield from remotely sensed water stress and solar radiation data. ISPRS J Photogramm Remote Sens 145:297–308.
[34] Van Ittersum M, Donatelli M (2003). Modelling cropping systems: Highlights of the symposium and preface to the special issues. Eur J Agron 18(3–4):187–197.
[35] Azzari G, Jain M, Lobell DB (2017). Towards fine resolution global maps of crop yields. Remote Sens Environ 202:129–141.
[36] Reichstein M et al. (2019). Deep learning and process understanding for data-driven Earth system science. Nature 566(7743):195–204.
[37] Taher F, Abdel-salam M, Elhoseny M, El-hasnony IM (2023). Reliable Machine Learning Model for IIoT Botnet Detection. IEEE Access 11:49319–49336.
[38] Khaki S, Wang L, Archontoulis SV (2020). A CNN-RNN framework for crop yield prediction. Front Plant Sci 10:1750.
[39] Kohavi R, John GH (1997). Wrappers for feature subset selection. Artif Intell 97(1–2):273–324.
[40] Talaat FM (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Computer Appl 35(23):17281–17292.
[41] Xing L, Li L, Gong J, Ren C, Liu J, Chen H (2018). Daily soil temperatures predictions for various climates in United States using data-driven model. Energy 160:430–440.
[42] United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. A/RES/70/1.
[43] Paudel D et al. (2023). Interpretability of deep learning models for crop yield forecasting. Comput Electron Agric 206:107663. *x*, pp. *xx–xx*, year.
[44] Kumar, Punith, and H. N. Champa. "Enhancing Agricultural Yield: A Unified Stacking Ensemble Method for Crop Recommendations using Soil Properties and Weather Attributes." *International Journal of Environmental Sciences* 11.6s (2025): 671-684.
[45] P. Kumar, H. Varun Prabhu and H. N. Champa, "Crop Recommendation using Ensemble Stacking Machine Learning approach," *2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon)*, HASSAN, India, 2023, pp. 1-6, doi: 10.1109/MysuruCon59703.2023.10396866.