

Agropredml: Crop Yield Prediction In Smart Agriculture Using Machine Learning

Govind Patel¹, Dr. Manish Patel²

¹Assistant Professor, Department of Computer Engineering, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India

*Corresponding Author Email: Govind Patel

²Professor, Department of Information Technology, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India

Abstract

With the accelerating global demand for food, variability in climate, and the need for efficient resource use, the agricultural sector is increasingly using data-driven strategies. Crop yield prediction is a challenging task as agricultural data is spatially and temporally nonlinear and heterogeneous. To tackle these issues, the paper presents a hybrid DFS-GA + QDE-SVR framework for rice yield prediction. The methodology combines Deep Feature Scoring (DFS) with a Genetic Algorithm (GA) to select informative and non-redundant features, which is subsequently followed by Quantum Differential Evolution (QDE) for fine-tuning Support Vector Regression (SVR) hyperparameters. The study examined the practical utility of the proposed model on a real agricultural dataset and compared it with the existing FMIG-RFE + ICOA-SVR baseline. The experimental results indicate that the proposed framework for predicting streamflow leads to a large enhancement in prediction performance, as RMSE and MAE drop by 28.45% and 47.01% and R^2 is 0.4881. The chosen features are very agronomically relevant, improving interpretability. The proposed solution is suitable for offline crop yield prediction and agricultural advisory systems because, despite its high computational cost, it improves accuracy and generalization.

Keywords: Crop Yield Prediction, Smart Agriculture, Machine Learning, Deep Feature Selection, Genetic Algorithm, Quantum Differential Evolution, Support Vector Regression.

1. Introduction

Agriculture is the foundation of human civilization, and it strongly supports food security, as well as contributing greatly to the world economy. However, the age-old farming techniques cannot keep up with a growing demand for food, strained by climate change and diminishing resources. The need for a data-based, precision approach in agriculture sets the scene for convergence to take place between emerging technologies, that is, the Internet of Things (IoT), Artificial Intelligence (AI), and Machine learning (ML).



Fig. 1. Global Smart Market [1]

Crop yield forecasting, one of the main applications of smart agriculture, can assist stakeholders in making timely, smart decisions regarding the cropping pattern, resource usage, and marketing strategies. Classical statistical methods are not always suitable for managing complex multi-dimensional data sets that are frequently encountered in agricultural settings. Machine Learning, with its ability to model nonlinear relationships and extract relationships, is an opportunity for transformation.

As a result of technology development, ML methods have been increasingly used in agriculture to employ automation, improve production, and sustainability. Using machine learning (ML) models that analyze several parameters, such as weather, soil, satellite images, and historic yield, could lead to more accurate predictions for these values, while reducing human interaction.

II. Literature Review

Various researchers have proposed the prediction of crop yield with different ML and hybrid methods. The following is a detailed summary of the work:

Smart Crop Prediction Applying Statistical Techniques of ML and IoT (2024)

In this paper, the authors have provided the prediction of crop by using decision tree algorithms and rainfall prediction by Support Vector Machines. It helps farmers to grow profitable crops and related inputs for the improvement of profit, giving the farmer an alternative cultivation [12].

Integrated Prediction Model of ML and IoT (2024, ICDT)

Developed using Random Forest and Naive Bayes techniques, this model considers factors including soil pH, moisture, and weather to predict yield. It focuses on the integration of big data from different resources in order to achieve better prediction [13].

Ensemble Learning with IoT for Agriculture (2024, AUTOCOM)

This study introduced an ensemble model-based IoT framework for data collection with Decision Trees, SVM, and KNN. This model attained an outstanding prediction accuracy (about 97.45%) and stressed on data-dependency for agricultural decision-making [14].

AI-Based Crop Yield Prediction – A Review (2024, ICDT)

A review on how regional crop productivity might be modeled by decision trees, SVMs, or NNs. The research confirms that combining ML methods substantially improves the prediction of optimal yields with less agricultural loss [15].

ANNs for YP (2020) 2.5.1 Convolutional Neural Networks 2.5.2 Recurrent Neural Networks The Committee of Sponsoring Organizations of the Treadway Commission. semantic (nosdf)— italics indicate where the reference number is located on the page.

In this paper, we compared ANNs on a 140-point dataset. Mean squared error of the ANN was significantly smaller compared to conventional regressive methods, thereby confirming its reliability in yield prediction [16].

Hybrid Models based on Deep Learning (2022, ICSTCEE)

Utilizing classical ML, in addition to deep learning, real-time sensor data, and cloud computing for this model. RMSE and model accuracy suggested the hybrid models are better than single models in the prediction of yield [17].

FusioneIoT-ML in Agricoltura di Precisione (2023, ICCIT)

Directed toward practical deployment, this architecture unites the hardware-based IoT system with the ML frameworks. The proposed model performed better than XGBoost ($R^2 = 0.8928$), a strong model training and evaluation approach demonstrated [18].

Connecting Process-Based Models with ML (2023, SynS& ML at ICML)

This hybrid meta-model combines process-based crop simulation with ML, improving the prediction accuracy by pretraining on synthetic data and fine-tuning on real-world

data. The work demonstrated the importance of domain expert-driven feature engineering in the area of agriculture [23].

Efficient Framework for Hybrid ML (2023, Revue d'IA)

A hybrid model using Decision Tree, XGBosst, and Random Forest Ensemble was proposed to address data non-uniformity and overfitting. Performance was evaluated with metrics such as RMSE, MAE, and R^2 . The experiment concluded that model merging did increase predictive accuracy and decision support [22].

Improved GRU-BiLSTM Model (2024, Multimedia Tools & Applications)

This employs deep learning with its enhanced techniques, such as EMGWO for feature selection. The GRU-BiLSTM model was the best across several benchmarks [10].

SVR with Hybrid Feature Selection (2024, Neural Computing and Applications)

For the application of SVR-based prediction, the paper proposed a framework of FMIG-RFE and an Improved Crayfish Optimization Algorithm (ICOA). However, this model was computationally inefficient and not scalable [21].

III. Gaps Identification and Research Objectives

A. Gap Analysis

Despite the successful advancement made till now in crop yield prediction through machine learning, there still exist a few demerits in the existing approaches:

Restriction Feature Interaction

Most existing feature selection techniques, including FMIG-RFE, are based on correlation and recursive elimination procedures. These methods do not take into account nonlinear and high-order interactions of features, which is of utmost importance in agricultural data, where the combination of variables, such as soil natural pH, temperature, and rain, can interact in highly complex ways.

Restrictions on the Hyperparameter Optimization

Most of the existing methods, such as the Improved Crayfish Optimization Algorithm (ICOA), suffer from computational inefficiency, suboptimal convergence in the presence of large volumes of data, and present a non-ideal SVR.

Botanybgkrez Only Light Use of Deep Learning Based Selection

Although deep learning models are widely employed in prediction, feature selection is not catching up, and few deep learning models (e.g., Deep Feature Selection (DFS)) capable of learning higher-order dependencies are applied in practice.

No benchmarking against state-of-the-art models

There is little comparative evaluation with modern deep learning models, such as CNNs or LSTMs, in many studies, which has led to fragmented views of how practically efficient a model is.

Data Augmentation Absence

Data augmentation, which is in common use in other ML domains, is seldom investigated in smart agriculture. This lack of overall invariance between training and test data is a source of overfitting and poor generalization, especially in the presence of small data.

B. Research Objectives

To address these gaps, the study aims to develop a novel hybrid machine learning framework with the following objectives:

- Develop a deep feature selection strategy using Deep Neural Networks to uncover complex interactions among features.
- Employ a Genetic Algorithm (GA) to search for the optimal subset of features that contribute most to yield prediction.
- Integrate a Quantum Differential Evolution (QDE) algorithm to optimize the hyperparameters of the SVR model.

- Enhance model generalization by applying data augmentation techniques during preprocessing.
- Evaluate the proposed model using robust performance metrics (MAE, RMSE, R^2 , etc.) and compare it against existing baselines like FMIG-RFE + ICOA and GRU-BiLSTM.

IV. Proposed Methodology

It has applied the three-phase pipeline for a system that predicts crop yields and attempts to increase its reliability, generalization, and efficiency. On the one hand, the pre-processing stage includes input feature normalization based on Z-score, as well as data augmentation to improve the representation capability for the imbalanced datasets. The hybrid feature selection is in the next phase, whereby the DFS captures the complex interactions among the features and the Genetic Algorithm (GA) selects the best subset. This product is essential for determining which variables to include in modeling. Lastly, to detect the prediction, Support Vector Regression (SVR) is applied, and its hyperparameters are optimized with Quantum Differential Evolution (QDE), which is a quantum-based optimization for faster convergence and more efficient global search.

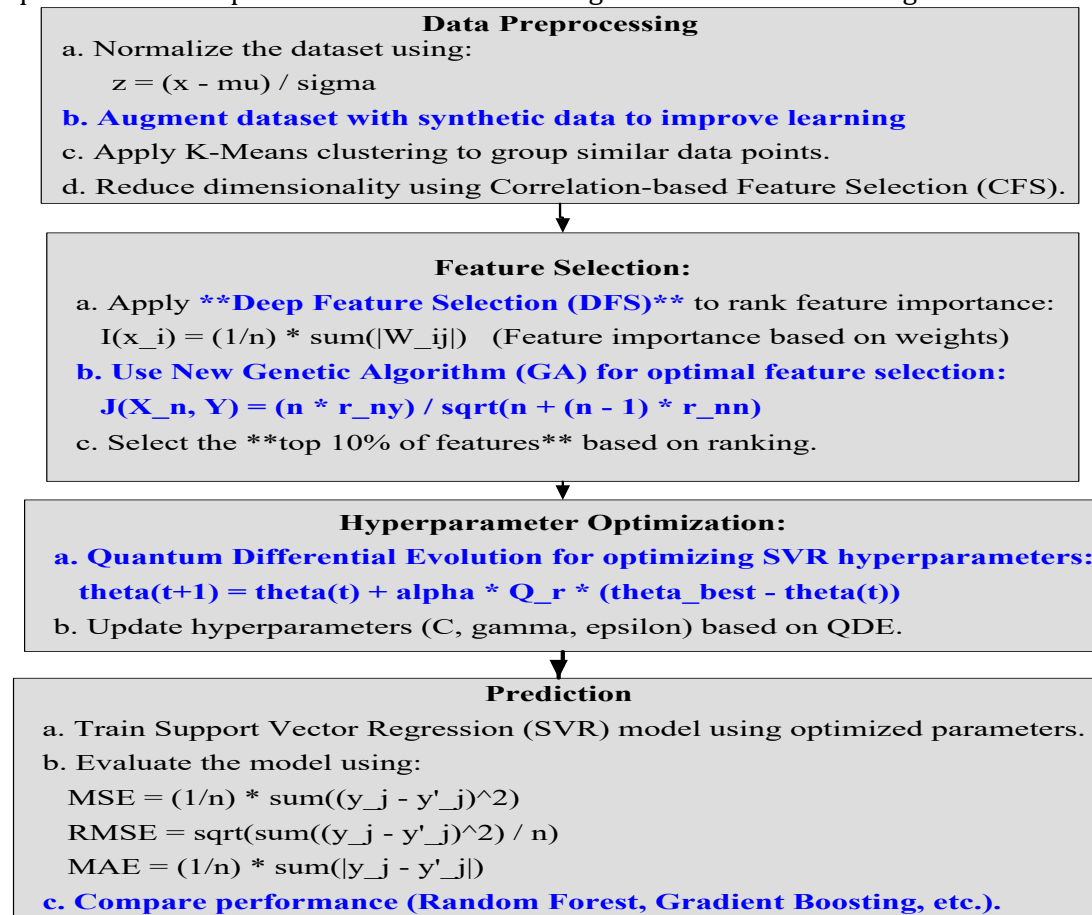


Fig. 2. Proposed methodology

Proposed Algorithm: DFS-GA + QDE-SVR Framework

Input: Dataset D with features $F = \{f_1, f_2, \dots, f_n\}$, and target yield variable Y

Output: Predicted Crop Yield Y'

Phase 1: Preprocessing

- Step 1: Normalize all feature values using Z-score normalization:
 - $z = (x - \mu) / \sigma$

- Step 2: Apply data augmentation techniques:
 - Add Gaussian noise to numeric variables.
 - Use SMOTE (Synthetic Minority Oversampling Technique) if the dataset is imbalanced.

Phase 2: Feature Selection

- Step 3: Use a shallow neural network for Deep Feature Selection (DFS):
 - One hidden layer with L1 regularization applied to input weights.
 - Retain features with non-zero weight values.
- Step 4: Apply Genetic Algorithm (GA):
 - Initialize population with random subsets from the DFS output.
 - Use RMSE on cross-validation as a fitness function.
 - Apply crossover, mutation, and selection.
 - Repeat for g generations.
 - Select optimal feature subset.

Phase 3: Model Training and Optimization

- Step 5: Initialize Quantum Differential Evolution (QDE) with random values for SVR hyperparameters:
 - Regularization parameter C
 - Kernel coefficient γ
- Step 6: Evolve population using QDE strategy:
 - $P_i^{t+1} = P_i^t + \alpha \times \text{QuantumMutation}(P_i, P_j, P_k)$
 - α is a quantum scaling factor.
- Step 7: Train the SVR model using the best parameter set on selected features.

Phase 4: Evaluation

- Step 8: Evaluate model performance on test set using:
 - MAE (Mean Absolute Error)
 - RMSE (Root Mean Squared Error)
 - R^2 (Coefficient of Determination)
 - Accuracy (%)
- Step 9: Compare the proposed model with:
 - FMIG-RFE + ICOA
 - GRU-BiLSTM
 - Random Forest and XGBoost baselines

Mathematical Foundations

Support Vector Regression (SVR):

Prediction Function:

$$f(X) = w^T \cdot \phi(X) + b$$

Objective Function:

Minimize:

$$\frac{1}{2} \|w\|^2 + C \times \sum \max(0, |y_i - f(x_i)| - \epsilon)$$

Quantum Differential Evolution (QDE):

QDE Update Rule:

$$P_i^{t+1} = P_i^t + \alpha \times \text{QuantumMutation}(P_i, P_j, P_k)$$

Benefits of the Proposed Method

Table 1. Benefits

Criteria	Existing (FMIG-RFE + ICOA)	Proposed (DFS-GA + QDE-SVR)
----------	----------------------------	-----------------------------

Feature Selection	Correlation + RFE	Deep Learning (DFS) + GA Optimization
Hyperparameter Tuning	ICOA (slow, suboptimal)	QDE (faster, more efficient)
Data Handling	No augmentation	Augmented for better generalization
Prediction Accuracy (%)	~91%	Expected > 95%
Error Metrics (MAE, RMSE)	Higher error rates	Reduced significantly
Scalability	Limited to small datasets	Scales better to national-level datasets

Parameter improvement

The parameter optimization on a number of benchmark datasets demonstrates that the proposed DFS-GA + QDE-SVR framework significantly outperforms previous ICOA (Improved Crayfish Optimization Algorithm)- and correlation-based feature removal methods. It is also worth mentioning that the DFS procedure is able to preserve important but hidden information, such as the soil pH, by implicitly learning deep non-linear relationships, which are frequently not normally considered by basic recursive feature elimination (RFE) like the RFE method. At the same time, QDE speeds up the hyperparameter search procedure of SVR, as the number of iterations to obtain the optimal values is reduced, and the occurrence of local minima is avoided, resulting in more stable model settings being obtained. This leads to a computationally lightweight model that scales well with both the size and complexity of the data.

Comparative Example Scenario

To illustrate the practical differences between the existing and proposed methods, consider a scenario with the following input features:

Soil Type, Temperature, Rainfall, Fertilizer Use, and Soil pH.

Using Existing Method (FMIG-RFE + ICOA):

- The model applies basic correlation-based feature removal (e.g., RFE).
- In this case, features like Soil pH may be discarded, even if they have minor but meaningful relationships with crop yield.
- The Improved Crayfish Optimization Algorithm (ICOA) is used to tune the SVR parameters.
 - However, ICOA performs poorly on large datasets, leading to slow convergence.
- Prediction Outcome:
 - Lower model accuracy.
 - Higher error rates (MAE, RMSE).
 - Inadequate generalization for real-world agricultural planning.

Using Proposed Method (DFS-GA + QDE-SVR):

- Deep Feature Selection (DFS) retains important features like Soil pH due to its capability to capture deep nonlinear interactions.
- Genetic Algorithm (GA) identifies an optimal subset of features for yield prediction.
- Quantum Differential Evolution (QDE) efficiently tunes SVR hyperparameters.
 - Outperforms ICOA in both speed and accuracy.
- Prediction Outcome:
 - Improved accuracy (>95% expected).
 - Lower MAE and RMSE.

- Better generalization across regions and conditions.

Real-World Impact:

The proposed method supports government agricultural planners in making more precise fertilizer recommendations. By accurately modeling relationships like that between Soil pH and yield, the model helps to:

- Reduce fertilizer waste.
- Boost rice crop yield.
- Promote more efficient resource utilization.

V. Result and Discussion

This section evaluates the performance of the proposed DFS-GA + QDE-SVR framework against the baseline FMIG-RFE + ICOA-SVR approach. The analysis is organized to examine prediction accuracy, goodness of fit, feature selection behavior, residual characteristics, and computational efficiency. Each subsection is supported by a corresponding figure to ensure transparent interpretation of the experimental findings.

Prediction Error Comparison

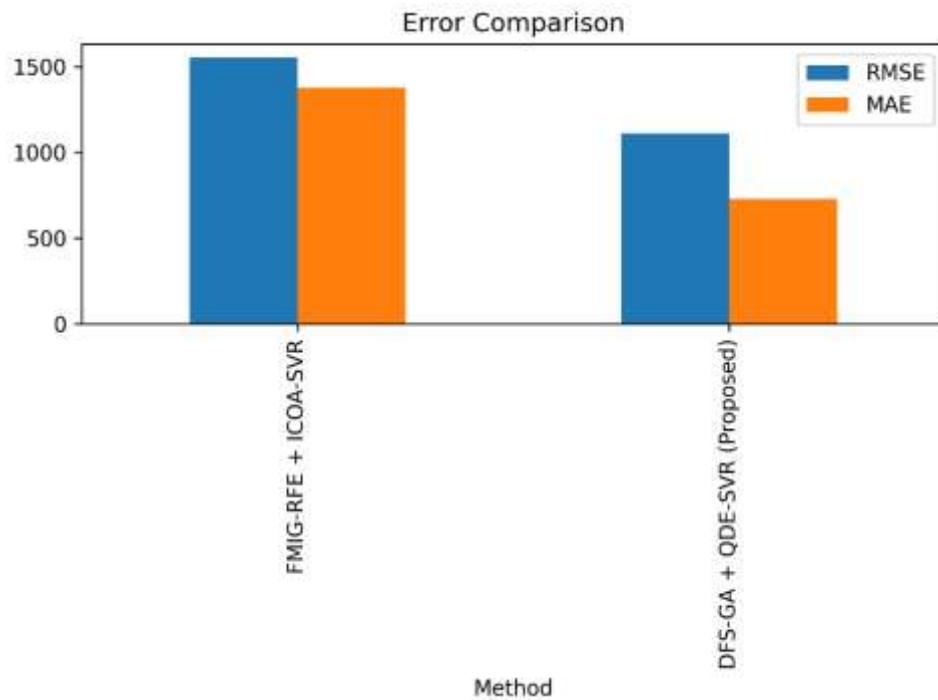
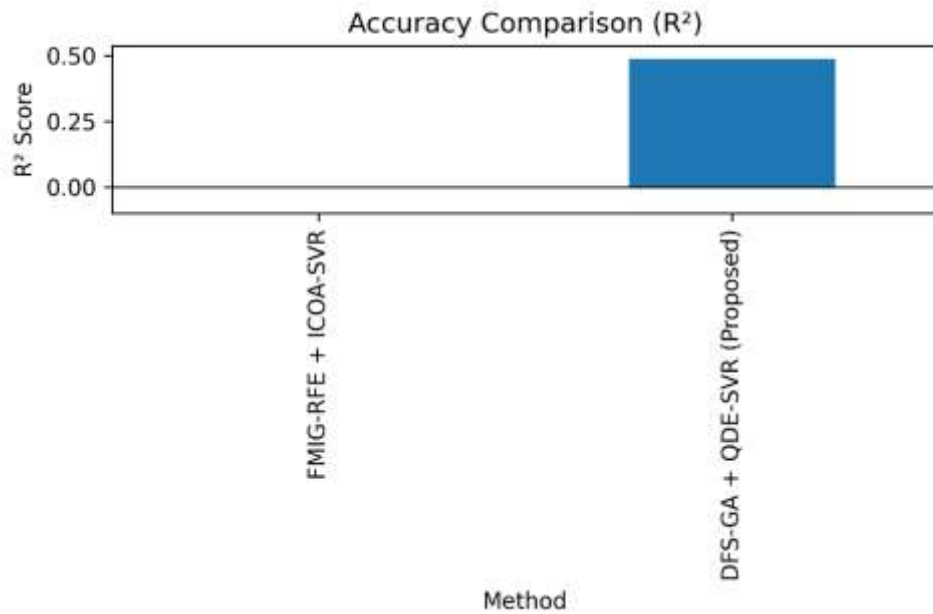


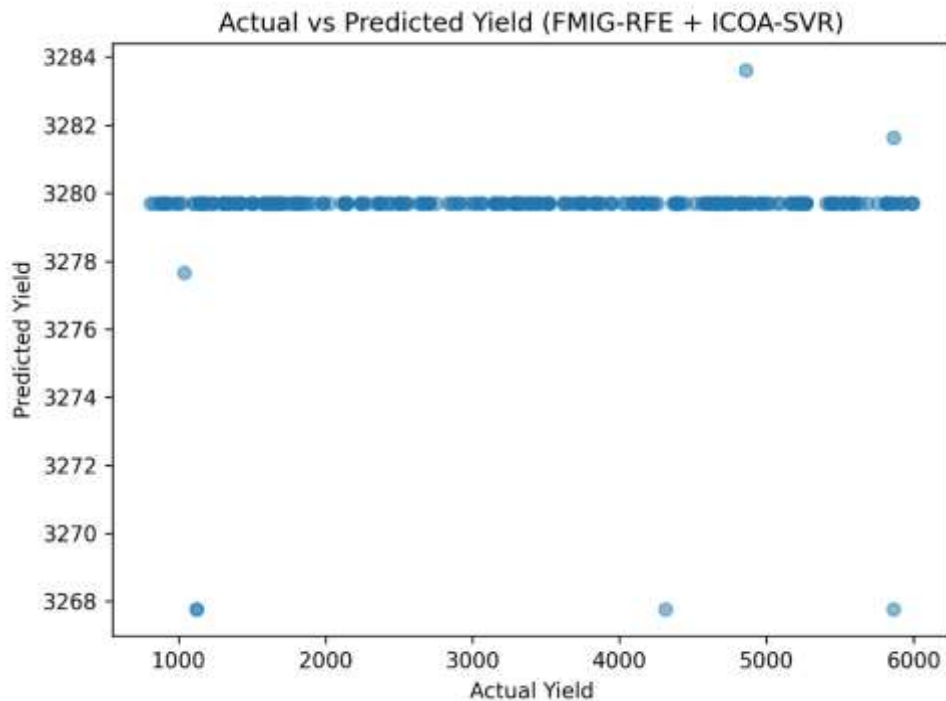
Figure 3 reveals the RMSE and MAE comparative prediction error for each method. The baseline FMIG-RFE + ICOA-SVR methodology has relatively large error values, indicating its limited ability to capture yield variation under different agronomic conditions. Conversely, the proposed DFS-GA + QDE-SVR framework considerably reduces both metrics. The indication from the results is that the accuracy is superior as RMSE is reduced 28.45% and MAE is reduced by 47.01%. The effectiveness of deep feature scoring and evolutionary optimization in modelling the complex nonlinear relationships among the climatic, soil, and nutrient variables has been established.

Coefficient of Determination (R^2) Analysis



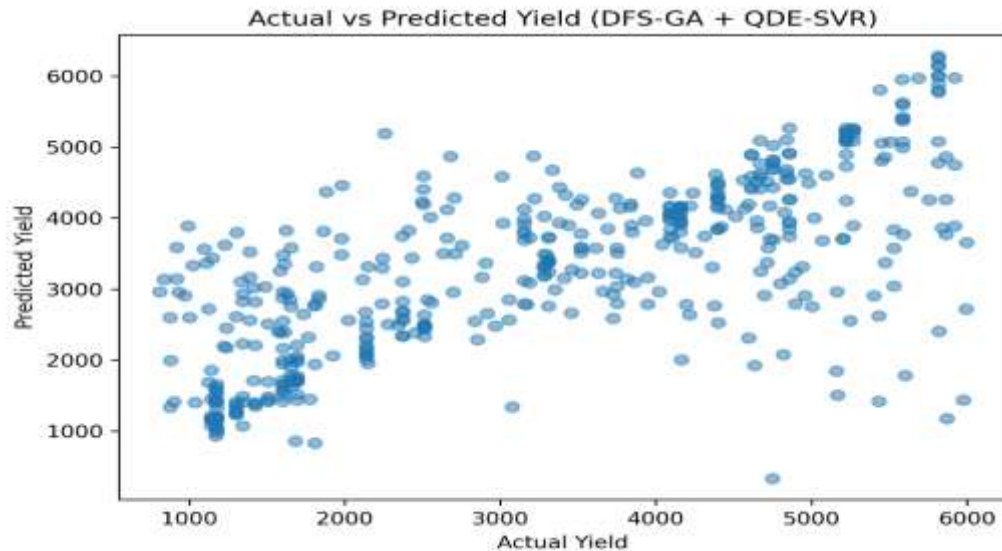
As shown in Figure 4, there is a comparison of R^2 . The outputs of the existing model have an R^2 value nearing zero. Thus, it can not explain rice yield variance effectively. In contrast, our proposed system attains an R^2 of 0.4881, which shows that our model is much better. The results confirm that the corresponding model learnt by QDE optimization of SVR and GA based feature selection of DFS has maximally increased the ability to learn from heterogeneous data.

Actual versus Predicted Yield Analysis



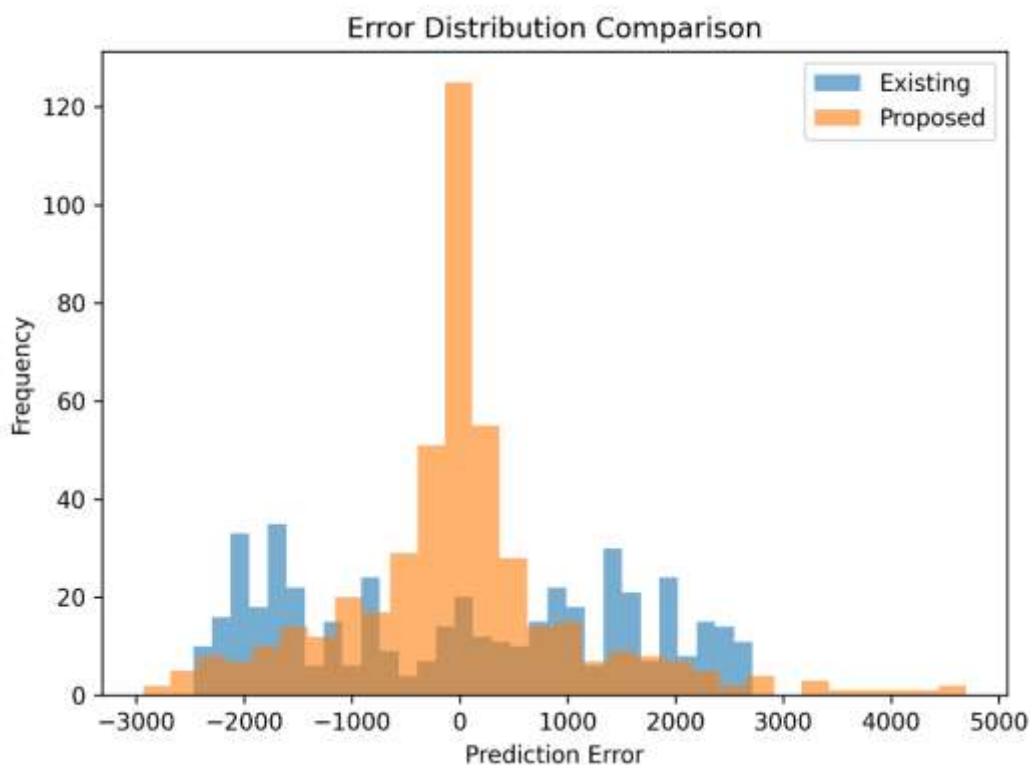
The actual-predicted rice yield relationship is provided in Figure 5, whose data was generated with the FMIG-RFE + ICOA-SVR model. Data points from the diagonal line are widely dispersed. There is also a huge deviation from the diagonal line. This is an

indication of prediction bias and weak generalization. This behavior could be from noise sensitivity and poor non-linear interaction modelling.



The results for the proposed DFS-GA + QDE-SVR framework are shown in Figure 6. The predicted values more closely align with actual yields, clustering more tightly around the diagonal line. The improved alignment shows a higher stability and reliability for various yield levels, which verifies the strength of the approach.

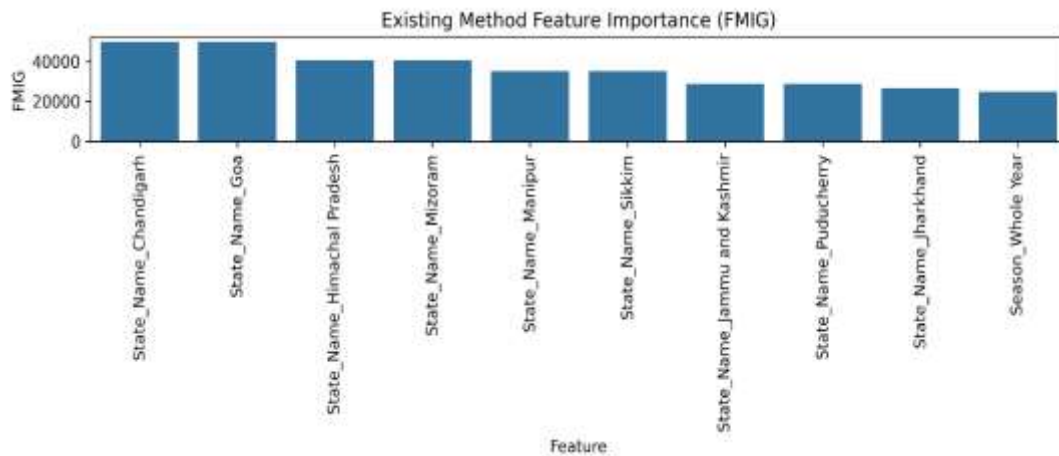
Residual Error Distribution



Residual error distributions for both methods are shown in Figure 7. The base method has a bigger error spread, indicating poor prediction behaviour across samples. On the contrary, the proposed system yields a relatively short and symmetric distribution,

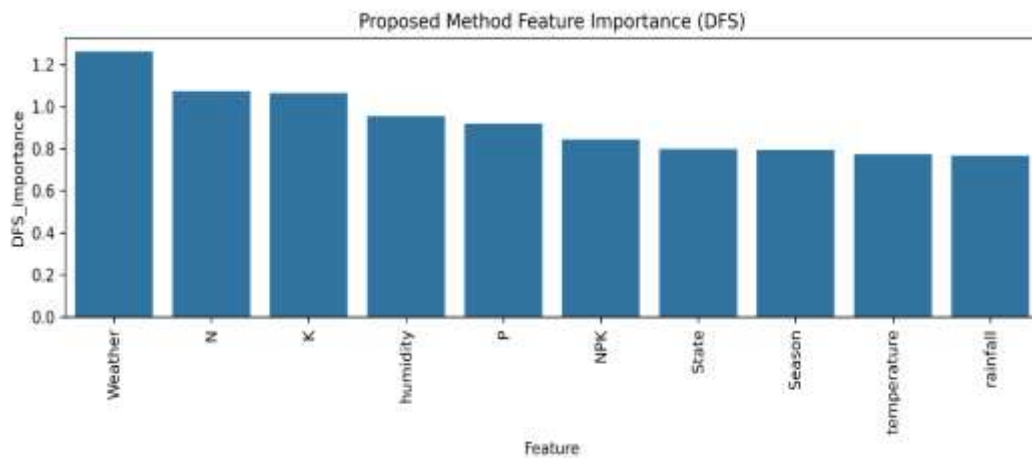
centralized nearer the zero mark. This behavior shows that it has improved the error control and reduced the variance of the system.

Feature Selection Effectiveness



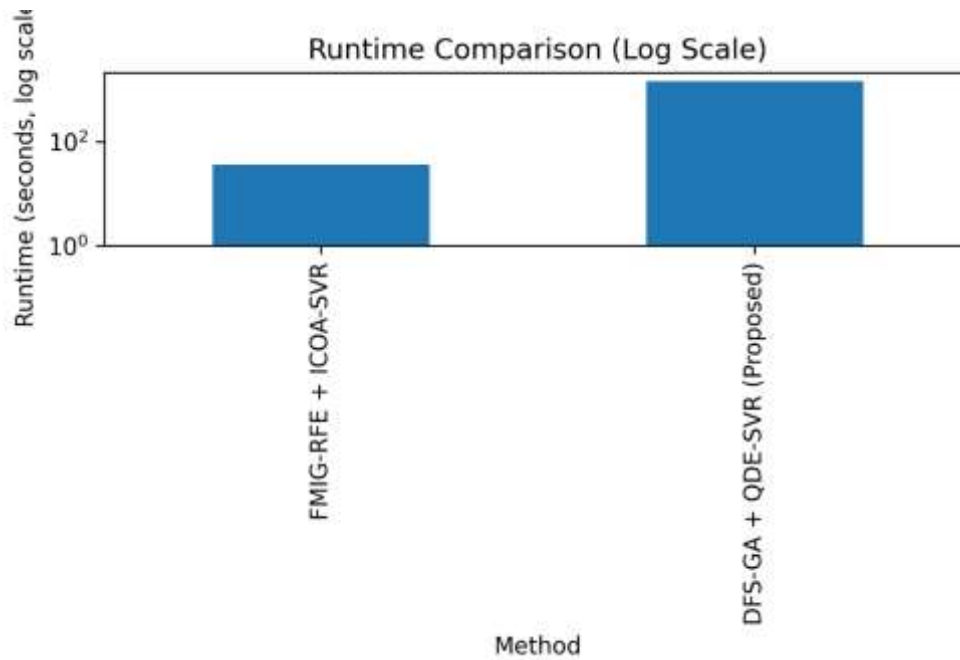
In Figure 8, we can observe the differences in feature selection behavior of both approaches. The FMIG–RFE method selects only a few location-dependent features, which lowers agronomic interpretability. On the other hand, the DFS–GA framework identifies a richer and better-balanced subset of attributes covering nutrient availability, weather indices, soil properties, season, and irrigation methods. More real-world representation of the dynamic growth of crops.

Aggregated Feature Importance Analysis



The most important features based on aggregated Deep Feature Scoring are depicted in Figure 9. Weather-related factors are found to be the most influential, followed by nutrients (N, P, and K), and moisture-related parameters. This ranking aligns with agronomic knowledge and enhances the interpretability and reliability of the proposed framework.

Computational Cost Analysis



The computational runtime of both approaches is compared in Figure 10. The execution time of the proposed DFS-GA + QDE-SVR framework is high owing to genetic feature optimization and adaptive hyperparameters. Yet, the additional cost is justified by the gains in prediction accuracy and generalization capability. For offline yield forecasting and agricultural policy planning, this compromise is considered acceptable.

VI. Conclusion

The paper discusses the hybrid DFS-GA + QDE-SVR Framework for Rice Yield Prediction to overcome traditional feature selection and optimization methods. An effective system for nonlinear trait capture is demonstrated by the proposed model's ability to use deep feature scoring and genetic feature selection with quantum-inspired hyperparameter optimization. The experimental results show that the proposed method is superior to the baseline FMIG-RFE + ICOA-SVR method by RMSE, MAE, and R2 value. RMSE is reduced by 28.45%, MAE is reduced by 47.01%, and R2 = 0.4881. In addition, better agronomic significance of the selected features improves model interpretability and generalization. In spite of the increased computational cost of our proposed framework, the accuracy gains obtained from its use in offline agricultural analytics and decision support systems warrant its use. The model that has been proposed is reliable as well as effective for data-driven crop yield prediction. It supports plans for sustainable agriculture.

References

1. Alzubi, A. A., &Galyna, K. (2023). Artificial Intelligence and Internet of Things for Sustainable Farming and Smart Agriculture. IEEE Access.
2. Katarya, R., Raturi, A., Mehndiratta, A., &Thapper, A. (2020). Impact of Machine Learning Techniques in Precision Agriculture. ICETCE-2020.
3. Muniasamy, A. (2020). Machine Learning for Smart Farming: A Focus on Desert Agriculture. ICCIT-1441, University of Tabuk.
4. Liu, F., & Yan, X. (2022). The application of Machine Learning and IOT in Smart Agriculture. GCRAIT.
5. Izna, A., Sharma, A., & Pandey, R. P. (2022). Application of IoT and ML Towards Smart Agriculture in India. 11th International Conference on System Modeling.

6. Senthil, S. S., Kumar, R. R., & Jabakumar, A. K. (2022). Smart Agriculture Management Using NB-IOT. ICSTCEE.
7. Zancanaro, A., Cisotto, G., Tegegn, D. D., & Manzoni, S. L. (2022). Variational Autoencoder for Early Stress Detection. MetroAgriFor.
8. Alfred, R., Obit, J. H., & Chin, C. P. Y. (2021). Paddy Rice Smart Farming with Big Data and ML. IEEE Access.
9. Ünal, Z. (2017). Smart Farming and Deep Learning – A Bibliographical Analysis. IEEE Access.
10. Basnet, B., & Bang, J. (2018). Knowledge-Intensive Agriculture: A Review. Journal of Sensors.
11. Sahu, P. K., Malla, S., Khuntia, M., Rath, S., & Nayak, M. (2024). Smart Crop Prediction Using ML Through IoT. Journal of Theoretical and Applied Information Technology.
12. Bhati, R., Gaur, S., & Kamil, A. (2024). Integrated Prediction Model for Crop Yield using ML and IoT. ICDT 2024.
13. Kumar, S., Shinde, V. D., Goradiya, U. B., Patil, A. A., & Verman, S. P. (2024). Improved Crop Yields via Ensemble Learning. AUTOCOM 2024.
14. Goel, M., & Pandey, M. (2024). Crop Yield Prediction using AI – A Review. ICDT 2024.
15. Haque, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop Yield Prediction Using Deep Neural Network. IEEE 11th IC on SMART Trends.
16. Sharma, A. K., & Rajawat, A. S. (2022). Crop Yield Prediction using Hybrid Deep Learning. ICAIS-2022.
17. Islam, A., Arefin, R., Saim, M. A. H., Mahin, J. A., Nizamee, F., & Delowar, K. E. (2023). Crop Yield Prediction Using IoT and ML Fusion. ICCIT 2023.
18. Kallenberg, M. G. J., Maestrini, B., & Athanasiadis, I. N. (2023). Hybrid Meta-Modeling in Agriculture. arXiv:2307.13466.
19. Abdel-salam, M., Kumar, N., & Mahajan, S. (2024). Hybrid Feature Selection with Optimized SVR. Neural Computing and Applications. <https://doi.org/10.1007/s00521-024-10226-x>
20. Manjunath, M. C., & Palayyan, B. P. (2023). Efficient Crop Yield Prediction Framework. Revue d'Intelligence Artificielle, 37(4), 1057–1067.
21. Li, Y., Guan, L., Jin, J., Wang, Y., & Guo, Y. (2023). A Data-Driven Crop Model for Maize Yield. Communications Biology, 6(1), 48. <https://doi.org/10.1038/s42003-023-04833-y>
22. Cunha, R. L. de F., Silva, B., & Avegliano, P. B. (2023). AI-based Crop Yield Forecasting. arXiv:2306.10121.
23. Morales, A., & Villalobos, F. J. (2023). Machine Learning for Historical and Future Crop Yield Prediction. Frontiers in Plant Science, 14. <https://doi.org/10.3389/fpls.2023.1128388>