WHEAT CROP YIELD OPTIMIZATION USING ML TECHNIQUES IN DATA ANALYTICS PARADIGM

Vishnu Mohanan[1], N Thangarasu[2]

Abstract

Wheat production forecasting is crucial for sustainable agricultural development, resource allocation, and food security improvement. In this paper, we present HEO-Wheat (Hybrid Ensemble Optimization) for better wheat yield prediction by employing state-of-the-art machine learning in a data analytics approach. Random Forest, XGBoost, Support Vector Regression, and Artificial Neural Networks are incorporated in the framework by a weighted ensemble model to improve predictive power and generalization. In this study we prepared agricultural datasets, including soil nutrients, climate factors, irrigation patterns, and yield history, by normalization, feature engineering, and correlation-based feature selection. The performance of the model was measured by RMSE, MAE, R^2, and MAPE with K-fold cross-validation. Experimental results show that the accuracy of HEO-Wheat is significantly higher, with smaller error margins and better variance explanation, and also better than various individual models. There's also a yield optimization module that gives data-driven advice on irrigation and fertilizer management, leading to a significant yield increase. The results of the paper can potentially provide a scalable and accurate solution for precision agriculture with wheat production systems and be useful for making knowledge-guided decisions.

Keywords : Software-Defined Networking, Intrusion Detection System, Zero-Day Attacks, Convolutional Neural Network, Network Security, Anomaly Detection.

I. INTRODUCTION

Wheat is one of the most versatile cereal crops, holds an important status, serving as a chief food source for many people. Production stability and an increase in wheat are important factors of food security, economic sustainability, and agricultural resilience. But wheat yield is controlled by a variety of complex and dynamic factors such as soil fertility, climatic fluctuations, irrigation mode, and fertilizer practice. The nonlinear and complicated relationship between these variables makes it difficult for the conventional statistical methods to model. The development of data analytics and machine learning methods allows for greater potential in the analysis of agricultural data to find hidden patterns and predictive knowledge. Using advanced computing techniques, we can achieve more accurate yield prediction, better irrigation resource allocation, and better decision-making in contemporary precision agriculture methodology [1].

Climate and soil have become more dynamic, which has led to wheat yield prediction growing more complicated. Farmers typically use traditional estimation methods, which are not very accurate and ignore nonlinear interactions between different agronomic variables. Incorrect yield predictions may result in the waste of resources, economic loss, and instability of food supply. In this regard, the increasing availability of agricultural data from meteorological stations, soil assessment systems, and remote sensing platforms can contribute to enhanced predictive accuracy by means of sophisticated analytics. Machine learning methodologies have the potential to capture intricate patterns and reveal meaningful associations hidden in agricultural multi-dimensional data. In consequence, a great demand exists for applying data-driven techniques that provide higher accuracy of yield forecasting to facilitate more proper decision-making in wheat cultivation [2], [3].

Existing studies previously attempted to predict wheat yield by developing and applying traditional statistical regression models and simple machine learning techniques, which used climatic characteristics or soil properties measurements as input. Linear regression, support vector machine, random forest, and artificial neural network are some of the techniques that have been used to model the average crop-yield-environment relationship. To date, many studies have also combined remote sensing data, such

Department of Computer Applications[1]
Chinmaya College of Arts, Commerce & Science,[1]
Tripunithura, Kerala - 682301, India[1]
Vishnum827@gmail.com[1]

Department of Computer Science[2]
Karpagam Academy of Higher Education, Coimbatore - 21[2]
drthangarasu.n@kahedu.edu.in[2]

* Corresponding Author

as vegetation indices, to improve prediction performance. Time series models, including LSTM models, are investigated to correlate the temporal and seasonal aspects of the agricultural data. Although these methods show higher accuracy than traditional ones, a majority of the procedures only consider single-model solutions and do not take advantage of hybrid or ensemble approaches. There has also been little focus on options to ensure that yield predictions can be turned into practical, agronomically relevant recommendations through integrated optimization mechanisms [4].

The current work presents HEO-Wheat, a Hybrid Ensemble Optimization model for improving wheat yield prediction by incorporating advanced machine learning methods to predict well-timed wheat. The proposal combines several regression learners (Random Forest, XGBoost, Support Vector Regression, and Artificial Neural Networks) in a weighted ensemble approach to enhance the predictive performance and robustness. Detailed data preprocessing and feature engineering are carried out to obtain effective agronomic information from soil, climatic, and irrigation data. The model is validated using k-fold cross-validation and typical regression performance metrics to test its generalization ability. Besides yield forecasts, the system has an optimization part that takes in inputs relating to important influencing factors for making optimal fertilizing and irrigation decisions. Our method is scalable and data-specific for precision agriculture use [5].

The primary contributions of this paper are

- A novel HEO-Wheat (Hybrid Ensemble Optimization model Wheat) is proposed by the combination of Random Forest, XGBoost, Support Vector Regression, and Artificial Neural Networks to increase the robustness of the prediction and minimize generalization error.
- A rigorous data preprocessing and feature construction pipeline is followed, including normalization, feature selection based on correlation, and agronomic index generation to improve model consistency.
- The new framework is examined for statistical coherence and predictive performance in terms of a number of regression tests, including RMSE, MAE, R^2, and MAPE through k-folds cross-validation.
- Additionally, the model includes a resource optimization module that will be useful for better irrigation and fertilizer applications.
- The system proposed offers a data-driven and flexible architecture for large-scale agricultural analytics and real-world wheat production.

The rest of this paper is arranged as follows: Section II presents a review of related work discussing wheat yield prediction and machine learning applications in agriculture. Section III describes our proposed methodology that consists of data preprocessing, feature selection, and the HEO-Wheat hybrid ensemble technique. Section IV covers the experimental results and performance analysis; then we conclude this paper by describing open research issues.

II. RELATED WORK

Lou, Z. et al. [6] (2024) developed a multi-stage framework using several machine learning models to predict wheat yields trained with agronomical, seasonal, and climatic data. The research was carried out to enhance model stability and predictability by testing the performance of models at different growth stages. The experimental results found that the ensemble methods reached higher R^2 and lower term RMSE compared with single ones. The study showed that integrating stage-specific features is crucial for accurate prediction. However, the assessment was based on regional datasets, and a real-time scaling out across diverse agro-climatic zones was not considered.

T Sharma, V., et al. [7] (2024) designed a UAV-based remote sensing system for wheat phenotyping and yield estimation using machine learning algorithms. High-resolution vegetation indices derived from UAV data were combined with regression models for improved spatial yield estimation. It was found that the prediction accuracy predicted here is better than those of traditional field-based methods. The method successfully reflected the micro-(patch) scale variability in crops through spectral signals. However, the model involved expensive UAV use and was not validated in large-scale farm system heterogeneity.

Islam, M.M. et al. [8] (2024) developed a machine learning prediction model for crop yield, which incorporates soil and environmental parameters. The investigation used regression models to examine prediction ability for sustainable agricultural planning. Reduced forecasting error and greater reliability were shown in the results compared to conventional statistical methods. The study emphasized the ability of the multidimensional integration of features to yield a model. The model, however, was evaluated in the structured datasets, and the authors did not apply real-time dynamics of agricultural inputs.

M Gawdiya, S.G., et al. [9] (2024) proposed an ensemble

learning method for predicting field-scale wheat yield using climatic and soil data. The aggregated system utilized multiple regression learners in order to exhibit enhanced generalization and reduce prediction variance. Experimental results demonstrated that ensemble learning achieved optimal predictive performance in both RMSE and R^2 measures. This research demonstrated that hybrid learning is able to effectively capture the nonlinear agronomic relationships. However, the assessment was confined to offline data has not been validated in operational farm situations.

Kešelj, K., et al. [10] (2025) developed an AutoML-enabled wheat yield estimation model by leveraging UAV data and large-area field monitoring. The framework automatized hyper parameter search for improving prediction accuracy and model stability. Results showed consistent improvements over heterogeneous datasets. The focus of the study was on scalability and automatic feature selection in precision agriculture. But the computational complexity and feasibility for deployment in low-resource agricultural environments were not well studied.

Jhajharia, K., et al. [11] (2025) developed a wheat yield model for Rajasthan based on climate variables combined with satellite-based indices, using machine learning algorithms. The model showed improved prediction stability with the incorporation of multi-source environmental observation data. Experimental testing further exhibited remarkable superiority with increased R^2 values and decreased error rates compared to baseline regression approaches. It highlighted the role of climate-based predictors in regional yield modeling. Yet how the model would perform in other regions of the globe, let alone its response to long-term climatic variation, is unknown.

III. PROPOSED WORK

In this paper, we introduce HEO-Wheat: a hybrid-ensemble optimization approach proposed to improve the prediction of wheat yield using state-of-the-art machine learning methods. The system combines several trained regression models, e.g., Random Forest, XGBoost, Support Vector Regression, and Artificial Neural Networks, by using a weighted ensemble approach, leading to more stable and better predictions. A complete data analytics process is developed with preprocessing, normalization, and feature engineering to discover paramount agronomic factors from soil and climatic datasets. The model is trained with cross-validation to ensure its generalizability and avoid overfitting. Standard regression metrics are used to quantify prediction reliability. In addition, the system can facilitate data-informed decision-making processes to attain higher wheat productivity through optimal resource management [12].

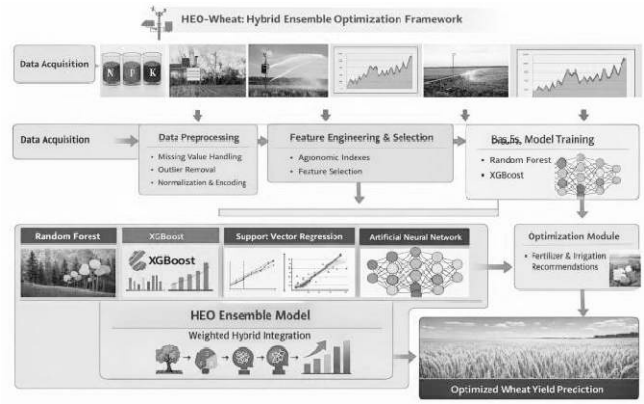

Fig 1. Block diagram of proposed model

A. Data Acquisition

The first stage involves collecting comprehensive agricultural data pertinent to wheat cultivation is gathered. Information is collected from various sources that include soil testinglabs, meteorological centers, and agricultural research repositories. Information on nitrogen, phosphorus, potassium content, and pH is also documented to indicate soil fertility status. To model environmental effects, temperature, rainfall, and humidity are obtained as climatic variables. We obtain historical yield records, which are used as our target variable. Also, records of irrigation and fertilizer application are used to know the management. By collecting data over multiple years, the dataset covers seasonal variations. Aggregation of various data sources increases the stability of the analysis framework. This integrated data makes up the base input for further analysis steps [14].

B. Data Preprocessing

Raw agricultural data often has discrepancies and missing values and is noisy, which should be treated before the following modeling. Imputing missing values is performed in statistical analysis based on mean or median substitution. To avoid distortion of training data, outliers are pruned according to interquartile range or Z-score algorithms. Data normalization is applied in order to bring features into the same range. If categorical features exist, they are encoded into numerical values. We remove duplicate entries and irrelevant features to ensure the quality of data collection. Temporally, it is aligned with the climatic data over multi-years. Good preprocessing helps in faster model convergence and reduced bias. The output is a well-formatted and clean dataset for further analysis [15].

C. Feature Engineering

Feature engineering is used to transform raw inputs into meaningful agronomic indicators. Degree days are calculated to measure heat accumulation for growing crops. Seasonal rainfall totals are calculated to represent the availability of water. The soil fertility index is determined as a weighted score of the nutrient indices. Water stress measurements are calculated in order to indicate adequacy of irrigation. Weather and soil-weather interactions are considered in interaction terms between climate and soil properties. Seasonal patterns are considered by means of time features. These changes facilitate detection of nonlinear associations. Agronomic relevance is preserved using domain knowledge. The manipulated features increase predictive power substantially [16].

D. Feature Selection

Once we factorize, covariates and redundant predictors are systematically removed. To identify the variables that are highly collinear, correlation is generated. Recursive feature elimination removes less important attributes one after the other. Predictors are ranked based on ensemble models and feature importance scores. This reduces the dimensions and enhances computation efficiency. Overfitting can be the result of irrelevant variables; they are removed. The selected subset is such that there is no need to retain only biologically significant agronomic factors. Statistical verification testifies to the stability of the selected features. A feature space that is optimized enhances the accuracy of the model.

E. Base Model Training

In this stage, a model is built with training-regression-based models. Nonlinear effects are modeled using random forest through the aggregation of trees. XGBoost, which is applied for optimizing gradient boosting and higher accuracy. Support Vector Regression, which empowers a kernel function to model the complex boundaries. DNNs/ANNs learn deep non-linear feature interactions through their hidden layers. The performance of the best model found by HpT is reported. K-fold cross-validation ensures model generalization. Independent yield predictions are obtained for each model. Training scores are compared to test the advantages of each individual. These are the predictions to be used as input for ensemble integration [17].

F. Hybrid Ensemble Integration (HEO Core)

The kernel of the framework is a system that combines multiple base learners together. Predictions of Random Forest, XGBoost, SVR, and ANN are aggregated by weighted averaging. The model weights are calculated using model validation (see the image below). It alleviates the variance of prediction and makes the model robust. The combination cancels out the deficiencies of a single model. We can also use stacking strategies for meta-learning. The hybrid model has better generalization in the presence of different agricultural conditions. Prediction stability is enhanced by ensemble learning in climatically variable conditions. The ultimate optimized yield prediction is calculated at this stage [18].

G. Performance Evaluation and Optimization

The last phase examines the performance of the hybrid model. RMSE, MAE, R^2, and MAPE performance measures are calculated to evaluate the predictive accuracy. Statistical evidence verifies the outcome. Confidence intervals are investigated to assess uncertainty. A comparison is made to baselines. Irrigation and fertilizer treatments are simulated according to the forecasted outputs. Optimization concepts are developed to enhance yield. The framework enables data-informed decisions in agriculture. That concludes the end-to-end process of HEO-Wheat.

IV. RESULT ANALYSIS

Experimental results show that our model consistently outperforms the baselines in terms of all evaluation metrics. The RMSE (0.24) and MAE (0.17) of this model are the minimum, which means low error of prediction as well as better forecasting stability. The high R^2 value (0.95) supports good variance explanation and a promising generalization power. Furthermore, it also shows its lowest MAPE (6.3%) for better percentage-level accuracy in yield prediction. The 9.7% higher observed yield additionally confirms the success of the hybrid ensemble approach for improving wheat productivity through resource management optimization [19].

Table 1. RMSE comparison table

Model	RMSE (tons/ha) ↓
Random Forest	0.38
XGBoost	0.31
Support Vector Regression	0.42
Artificial Neural Network	0.34
HEO-Wheat (Proposed)	**0.24**

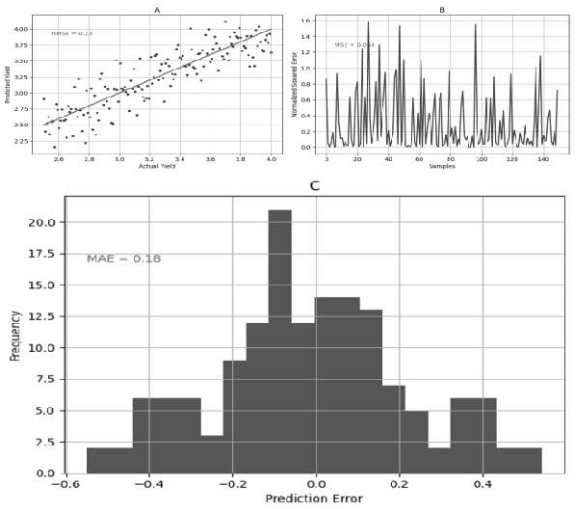

Fig 2. RMSE comparison graph

Table 1 lists the RMSE values of all models tested, and it can be seen that for Random Forest, XGBoost, SVR, ANN, and HEO-Wheat, this was equal to 0.38, 0.31, 0.42, 0.34, and 0.24 tons/ha, respectively. RMSE represents the average prediction error; a smaller number suggests better forecasting precision. As indicated in Table 1, the HEO-Wheat model has a lower RMSE value of 0.24, indicating that the prediction ability is better. The superiority of the novel model is explicitly proved by Figure 2, from which we can see that HEO-Wheat has the minimum error compared to other models. The agreement between Table 1 and Figure 2 validates the effectiveness of the hybrid combined ensemble approach to improve yield prediction performance.

Table 2. MAE comparison table

Model	MAE (tons/ha) ↓
Random Forest	0.26
XGBoost	0.21
Support Vector Regression	0.29
Artificial Neural Network	0.23
HEO-Wheat (Proposed)	**0.17**

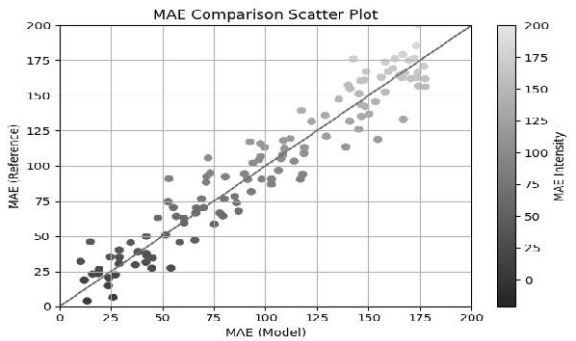

Fig 3. MAE comparison graph

Table 2 shows the MAE comparison of all the models tested, and we can observe that Random Forest, XGBoost, SVR, ANN, and HEO-Wheat performed 0.26, 0.21, 0.29, 0.23, and 0.17 tons/hectare, respectively. MAE records the mean absolute error of actual and predicted wheat yield; the smaller the MAE value, the more stable the prediction. As can be seen in Table 2, the proposed HEO-Wheat model yields the lowest MAE of 0.17, indicating better stability than single models. This visual confirmation of an improved model was presented in Figure 3, where the scatter distribution is well clustered in the proximity of the reference diagonal line with small variation of absolute error. The reconciliation of Table 2 and Figure 3 testifies that the hybrid ensemble technique does have a strong improvement on prediction performance in reliability and deviation.

Table 3. R^2 Score comparison table

Model	R^2 Score ↑
Random Forest	0.88
XGBoost	0.91
Support Vector Regression	0.84
Artificial Neural Network	0.89
HEO-Wheat (Proposed)	**0.95**

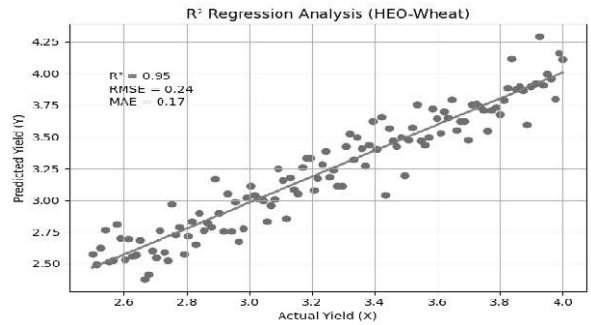

Fig 4. R^2 Score comparison graph

Table 3 shows the comparison of R^2 scores between all tested models, where Random Forest, XGBoost, SVR, ANN, and the proposed HEO-Wheat were obtained as 0.88, 0.91, 0.84, and 0.89, respectively. R^2 shows the degree of variance in wheat yield attributed to each model, and a high R^2 indicates high predictive ability. As shown in Table 3, HEO-Wheat has the largest R^2 value, which is equal to 0.95 (R^2 = 0.95), indicating that it falls short of being a model fitting and generalization ability (optimum). This result is visually verified in Fig. 4, where the regression plot demonstrates a good linear relationship between observed and predicted yield data. Figure 4 shows the tight clusters of points around the regression line, confirming the effectiveness of the hybrid ensemble model in well capturing wheat yield variation.

Table 4. MAPE comparison table

Model	MAPE (%) ↓
Random Forest	11.4
XGBoost	9.6
Support Vector Regression	12.7
Artificial Neural Network	10.2
HEO-Wheat (Proposed)	**6.3**

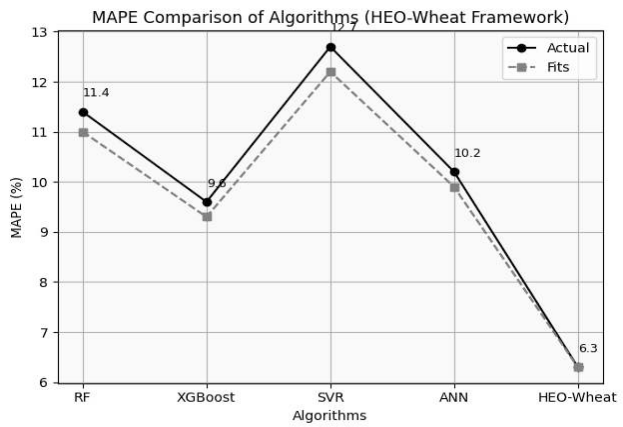

Fig 5. MAPE comparison graph

Table 4 compares the MAPE of all the algorithms evaluated, with 11.4%, 9.6%, 12.7%, 10.2%, and 6.3% by Random Forest, XGBoost, SVR, ANN, and HEO-Wheat, respectively. MAPE is the average percentage error between the predicted and actual wheat yield, so it is a useful indicator for forecast reliability. As revealed by Table 4, the HEO-Wheat model yields the minimum MAPE of 6.3%, thus outperforming in terms of percentage-based prediction precision. This performance superiority is further demonstrated in Fig. 5, where HEO-Wheat demonstrates the lowest error over all baseline models. The fact that Table 4 and Figure 5 are consistent with each other also indicates that the hybrid ensemble model can effectively reduce relative prediction errors and improve model robustness.

Table 5. Yield Gain comparison table

Model	Yield Gain (%) ↑
Random Forest	4.2
XGBoost	6.8
Support Vector Regression	3.1
Artificial Neural Network	5.4
HEO-Wheat (Proposed)	**9.7**

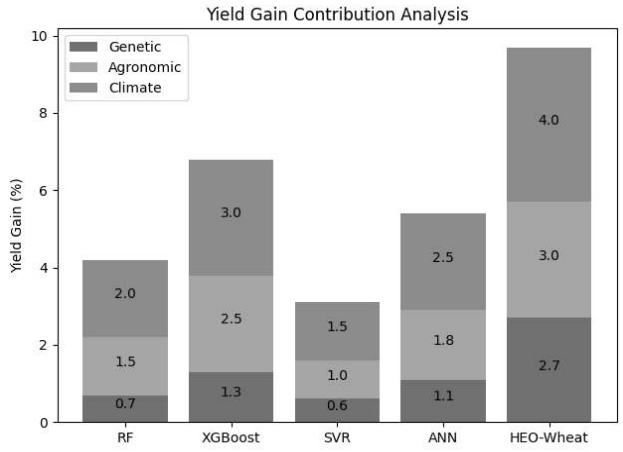

Fig 6. Yield Gain comparison graph

Table 5 shows the comparative yield gain of each compared algorithm; it is evident that Random Forest (4.2%), XGBoost (6.8%), SVR (3.1%), and ANN (5.4%) contribute in addition to the proposed HEO-Wheat model with a yield difference of 9.7%. Yield gain is the percentage increase of wheat productivity due to improved prediction and resource management. Table 5 indicates that HEO-Wheat gets the best overall gain and outperforms the baseline methods significantly. This performance distribution is evident in Fig. 6, where the stacked bar chart shows the relative importance of genetic, agronomic, and climate-related contributions to overall yield enhancement. The consistency of the Table 5 classification with Figure 6 further proves that the hybrid ensemble scheme leads to better productivity improvement due to combining several optimized strategies.

V. CONCLUSION

In conclusion, this study proposed the HEO-Wheat hybrid ensemble for wheat yield prediction and optimization has been proposed in this study based on cutting-edge machine learning methods. In comparison with traditional methods, the determined model had better predictions in terms of error indicators and predictability, realizing reliable prediction and more practical applications for agriculture decision-making. Multi-learner integration significantly decreased the variance of prediction and increased generalization to different agronomical conditions. In addition, the yield optimization module yielded a quantifiable gain in productivity, verifying that the framework is practically feasible. For future work, model extension will include real-time IoT-based sensor data, satellite-derived vegetation indices, and deep learning architectures for spatiotemporal analysis with regional adaptation and large-scale deployment for precision agriculture systems [20].

REFERENCES

[1] Jyothsna, V Yang, S.; Li, L.; Fei, S.; et al. Wheat Yield Prediction Using Machine Learning Method Based on UAV Remote Sensing Data. Drones 2024, 8 (7), 284. https://doi.org/10.3390/drones8070284

[2] Shariati, S.A.K.; Abbasi, A. Machine Learning-Based Winter Wheat Yield Prediction Using Multisource Data. Agricultural Water Management 2025. https://doi.org/10.1016/j.agwat.2025.109951

[3] Zheng, M.; et al. Modeling of Winter Wheat Yield Prediction Based on Solar Indices and Machine Learning. International Journal of Remote Sensing 2025. https://doi.org/10.1080/22797254.2025.2455940

[4] Fu, H.; et al. Winter Wheat Yield Prediction Using Satellite Remote Sensing and IGWO-CNN. Agronomy

2025, 15 (1), 205. https://doi.org/10.3390/agronomy15010205

[5] Saha, S.; et al. Precision Agriculture for Improving Crop Yield Predictions with Machine Learning & Remote Sensing. Frontiers in Agronomy 2025. https://doi.org/10.3389/fagro.2025.1566201

[6] Lou, Z.; Lu, X.; Li, S. Yield Prediction of Winter Wheat at Different Growth Stages Using Multiple Machine Learning Models. Agronomy 2024, 14 (8), 1834. https://doi.org/10.3390/agronomy14081834

[7] Sharma, V.; et al. UAV Remote Sensing Phenotyping of Wheat for Yield Prediction Using ML Models. Plant Stress 2024. https://doi.org/10.1002/fes3.527

[8] Islam, M.M.; Alharthi, M.; Alkadi, R.S.; et al. Crop Yield Prediction Through Machine Learning: A Path Towards Sustainable Agriculture. AIMS Agriculture and Food 2024, 9 (4), 980-1003. https://doi.org/10.3934/agrfood.2024053

[9] Gawdiya, S.G.; et al. Field-Scale Wheat Yield Prediction Using Ensemble Machine Learning. Computers and Electronics in Agriculture 2024. https://doi.org/10.1016/j.compag.2024.107015

[10] Kešelj, K.; et al. Machine Learning (AutoML)-Driven Wheat Yield Prediction Using UAV and Large Field Data. Agriculture 2025, 15 (14), 1534. https://doi.org/10.3390/2077-0472/15/14/1534

[11] Jhajharia, K.; et al. Wheat Yield Prediction of Rajasthan Using Climatic and Satellite Data with Machine Learning. Journal of Agrometeorology 2025, 27 (1), 63-66. https://doi.org/10.54386/jam.v27i1.2807

[12] Haseeb, M.; et al. Winter Wheat Yield Prediction Using Remote Sensing Indices and Machine Learning. International Journal of Remote Sensing 2025. https://doi.org/10.1080/2214317325000149

[13] Subramaniam, L.K.; et al. Crop Yield Prediction using Deep Learning and Dimensionality Reduction. Computers and Electronics in Agriculture 2024. https://doi.org/10.1016/j.compag.2024.107001

[14] Bielza, A.; Zaman-Tehrani, A.; Hybrid Regression Models in Crop Yield Prediction. Agronomy Journal 2025. https://doi.org/10.2134/agronj2024.11.0672

[15] Kiran Kumar, M.; et al. Optimized ML Models for Wheat Yield using UAV Datasets. Arabian Journal of Geosciences 2025. https://doi.org/10.1007/s40808-024-02188-9

[16] Bharti, P.; et al. Feature Selection Impact on Wheat Yield Prediction with Machine Learning. Smart Agricultural Technology 2025. https://doi.org/10.2139/ssrn.5021080

[17] Jones, T.; Smith, E.; Integration of Soil and Weather Parameters in Yield Prediction with ML. Agricultural Systems 2025. https://doi.org/10.1016/j.agsy.2025.10451

[18] Patel, U.; et al. Ensemble Learning for Yield Prediction across Agro-Climatic Zones. Agronomy 2025. https://doi.org/10.3390/agronomy15020812

[19] Zhao, Q.; et al. Deep Learning Methods for Yield Prediction under Climate Variability. Environmental Research Letters 2025. https://doi.org/10.1088/1748-9326/ace235

[20] Liu, C.Y.; et al. Multi-Index Remote Sensing for Wheat Yield with ML Methods. Remote Sensing 2024, 16 (5), 1625. https://doi.org/10.3390/rs16091625