Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 429–440
DOI: 10.15439/2025F5652
ISSN 2300-5963 ACSIS, Vol. 43

Multi-level Optimization-based Ensemble Machine Learning for Efficient Crop Yield Prediction in Saudi Arabia

Raya Aldawoud*, Zohra Sbaï*†
*Department of Computer Science, Prince Sattam bin Abdulaziz University, Al Kharj, Saudi Arabia
†National Engineering School of Tunis, Tunis El Manar University, Tunis, Tunisia

z.sbai@psau.edu.sa

***Abstract*—Accurate crop yield prediction is essential for enhancing food security, optimizing resource use, and supporting smart farming initiatives. Traditional statistical models often fail to capture the nonlinear interactions between environmental, climatic, and agronomic variables. To address these challenges, this research evaluated seven ensemble machine learning models: Random Forest, Bagged Decision Tree, AdaBoost, Gradient Boosting, Stochastic Gradient Boosting, eXtreme Gradient Boosting, and CatBoost. Being interested in the region of Saudi Arabia, we collected and integrated multi-source data that resulted in meteorological factors (temperature, humidity, precipitation), vegetation indices, pesticide usage, and historical yield records of thirty eight crop categories. While selecting only eight crops to study in the optimization phase, each model was assessed under four experimental configurations: baseline models, hyperparameter tuning, outliers' removal, and Bayesian optimization. Results showed that the optimized models performed up to 47% better than the default models. More precisely, hyperparameter tuning showed marginal gains, and Bayesian optimization and outliers' removal led to noticeable performance improvements. However, the effectiveness of each strategy was crop-specific.**

***Index Terms*—Crop Piield Prediction, Ensemble Machine Learning, Bagging, Boosting, Hyperparameter Tuning, Bayesian Optimization**

I. Introduction

ENSURING sustainable food production is a global challenge, especially in arid regions like Saudi Arabia, where agriculture is constrained by extreme climate, water scarcity, and limited arable land. Crop yield prediction has emerged as a critical tool to enhance agricultural planning, reduce resource waste, and support national food security strategies. Accurate forecasting allows for informed decisions in land management, irrigation scheduling, and fertilizer application, while also aiding in mitigating the risks of climate variability and supply chain disruptions.

Despite the advances in machine learning (ML) and deep learning (DL) methods for yield prediction, many existing studies are either based on synthetic or global datasets and rarely reflect the local agro-climatic and environmental realities of the Gulf region in general. Moreover, predictive models are often developed for major cereal crops like wheat and maize, with limited attention to a broader range of fruits and vegetables that are vital to the regional diet and agricultural economy. The present study addresses these gaps by developing an ensemble learning-based framework to forecast the yield of several crops grown in Saudi Arabia.

We propose a framework that leverages multiple ensemble learning techniques including Random Forest, Bagging Decision Trees, Gradient Boosting, Stochastic Gradient Boosting, XGBoost, AdaBoost, and CatBoost to exploit the strengths of individual models while enhancing robustness and accuracy. For each crop, the models are trained and evaluated using historical yield data alongside environmental variables such as temperature, precipitation, humidity, NDVI, VCI, WDRVI, and pesticide application rates.

In our research, we seek to develop predictive models specifically tailored to key crops grown in Saudi Arabia by utilizing real, localized agricultural data. The study aims to systematically compare the performance of various ensemble learning algorithms across multiple crop types to determine their relative strengths and limitations. Also, it seeks to identify the most effective model for each crop by evaluating their performance using standard regression metrics such as the coefficient of determination (R^2), Mean Absolute Error (MAE),Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Ultimately, the research aims to provide a scalable and reliable modeling framework that supports yield forecasting, enhances strategic food security planning, and promotes the adoption of smart agriculture practices in arid and data-constrained environments.

By combining local data with advanced ensemble learning strategies, this research contributes to the growing body of precision agriculture literature and offers practical insights for decision-makers aiming to modernize agriculture in Saudi Arabia. More precisely, the contributions of the present research are the following:

- Unified ensemble-based prediction framework tailored to the Saudi agro-environmental context.
- Comprehensive comparative analysis of seven ensemble machine learning models across multiple crop categories, revealing insights into their relative performance under different optimization conditions.

- Multi-stage optimization pipeline to enhance models' accuracy and stability; including direct optimization techniques such as hyperparameter tuning and Bayesian search and methods of enhancing performance like outliers' removal.

The remainder of the paper is organized as follows. Section 2 recalls the works existing in the literature for crop yield forecasting. Section 3 details the methodology adopted to realize the aforementioned contributions. The main results are discussed in section 4. Finally, section 5 concludes the paper.

II. Related work

Crop yield prediction has emerged as a critical area of research in agriculture [32], especially in the face of climate change, population growth, and food security challenges. Accurate prediction of agricultural production enables farmers, and agribusiness stakeholders to make decisions about resource allocation, land management, and market planning. It also plays a role in supporting food sustainability and reducing risks in agricultural investment and planning [11], [18].

In recent years, artificial intelligence (AI) and machine learning (ML) have been increasingly adopted for crop yield forecasting due to their capacity to model complex, nonlinear relationships among environmental, soil, and agronomic variables [12], [18]. These technologies have shown results compared to traditional statistical methods, especially when dealing with large datasets.

Palanivel and Surianarayanan [13] reviewed ML methods for crop yield prediction, highlighting ANN and SVM for handling nonlinear agricultural data. Studies from India, China, and South Africa covered wheat, maize, rice, cotton, and potatoes. Key features: climate (rainfall, temperature, humidity) and soil. ANN showed high accuracy—e.g., $\pm 9\%$ error in wheat yield. A big data framework was proposed to improve prediction via distributed processing.

Ashfaq et al. [1] proposed an ML framework for winter wheat yield prediction in Multan, Pakistan (2017–2022), integrating meteorological, NDVI, soil, and spatial data. Three models—SVM, RF, and LASSO—were evaluated using NDVI (Landsat 8), climate, and soil features. RF outperformed others with $R^2 = 0.88$, 97% accuracy, and RMSE of 0.056 t/ha. NDVI and climate data boosted accuracy, with water-related features more impactful. The study also produced spatial yield maps and highlighted RF's suitability for heterogeneous, data-scarce regions.

Ilyas et al. [4] proposed an AI-based framework for automated crop classification and yield prediction using remote sensing and ensemble learning. A fuzzy hybrid ensembled model was developed, combining spatial filtering and data augmentation with a bagging ensemble classifier. Yield prediction used GB, RF, DT, and linear regression trained on FAO and World Bank data (2017–2021) for flaxseed, lentils, rice, sugarcane, and wheat. The ensemble classifier improved accuracy by +13% over GB and +24% over DT. GB regressor achieved the best performance with lowest MSE and highest accuracy.

Yewlea et al. [2] proposed RicEns-Net, a deep ensemble model for rice yield prediction in Vietnam's An Giang province using multi-modal data: Sentinel-1 (SAR), Sentinel-2 (MSI), Sentinel-3 (meteorology), NASA rainfall, and field data. From 557 observations, 15 features were selected via statistical filtering. RicEns-Net combines CNN, MLP, DenseNet, and AE, weighted by validation error. It outperformed all tested models, achieving MAE = 341.13 kg/ha, RMSE = 436.26 kg/ha, and adjusted R^2 = 0.589, with minimal train-test variance (ΔR^2 = 0.063).

Wang et al. [6] reviewed DL models for yield prediction across crops: corn, soybean, rice, wheat, tomato, and lettuce. They evaluated LSTM, CNN, CNN-RNN, BO-LSTM, ConvLSTM, DNN, and GRU against ML models (RF, SVR, GBR, KNN), using meteorological, soil, remote sensing, and management data. Notable results: BO-LSTM ($R^2 = 0.82$ for wheat), GRU ($R^2 = 0.98$ for corn), CNN-RNN ($R^2 = 0.9995$ for tomato). DL outperformed ML on large datasets but at higher computational cost and lower interpretability.

Morales and Villalobos [15] evaluated ML models for wheat and sunflower yield prediction using synthetic data from DSSAT models across five Spanish regions. Features included weather, soil, cultivar, and management practices. Models tested: Lasso, Ridge, RF, and ANN (ANN-2 to ANN-12). RF achieved best test results for both crops (e.g., wheat: RMSE = 5%, R^2 = 0.99 train; sunflower: RMSE = 12%, R^2 = 0.97). ANN models showed overfitting risks. Chronological validation confirmed RF's robustness over random splits, supporting its use in real-world forecasting.

El-Kenawy et al. [21] evaluated ML and DL models for potato yield prediction using a Kaggle dataset with agro-climatic, temporal, and spatial variables. Models included KNN, GB, XGBoost, MLP (ML), and GNN, GRU, LSTM (DL). GNN outperformed all, achieving $R^2 = 0.5172$ and MSE=0.0236. LSTM and GRU followed with $R^2 = 0.4474$ and 0.3872. DL models better captured spatial-temporal dependencies. The study applied extensive preprocessing, hyperparameter tuning, and emphasized DL's value for sustainable agriculture and food security.

Joshi et al. [3] explored deep transfer learning for winter wheat yield prediction in U.S. climate zones using satellite time-series and climate data (2008–2020). A BiLSTM replaced MLP to better model spatiotemporal features. Four transfer techniques were tested: TrAdaBoost.R2, Two-stage TrAdaBoost.R2, DANN, and Fine-tuning. Two-stage TrAdaBoost.R2 + BiLSTM yielded top results (MAE = 0.41/0.42; R^2 = 0.51/0.53). DANN underperformed due to domain mismatch. The study highlights BiLSTM's and domain similarity's roles in enhancing transfer learning efficacy.

Brandt et al. [5] developed an ensemble learning framework for predicting yields of winter wheat, barley, and rapeseed in Germany using data from 140,000–155,000 parcels (2019–2022). Features included Sentinel-2 imagery, meteorological data, and soil properties. Two ensemble strategies—stacking (PLSR, RF, SVR, CTB, LGB, XGBoost with Elastic Net meta-learner) and majority voting—were tested.

The voting ensemble achieved the best parcel-level R^2 scores: wheat (0.74), barley (0.68), rapeseed (0.66). High-resolution EO and agro-climatic fusion enhanced accuracy and scalability.

Rao et al. [11] proposed a supervised ML model to guide crop selection based on soil nutrients (N, P, K), temperature, humidity, pH, and rainfall, using a dataset of 2200 samples across 22 crops. Evaluated models included KNN, DT, and RF classifiers. RF achieved top test accuracy (99.32%) with both Gini and Entropy. DT reached up to 98.86%, while KNN scored 97.04%. Regression analysis (ENet, Lasso, Kernel Ridge) and stacking were also used for yield approximation. A mobile app with GPS/rainfall tracking was suggested for practical deployment.

In Saudi Arabia, multiple research papers have studied crop yield prediction. For instance, Assous et al. [12] developed a neural network model for yield prediction across Gulf countries using some features such as temperature, rainfall, nitrogen, and pesticide use, achieved a determination coefficient (R^2) of 0.93. Likewise, Al-Adhaileh and Aldhyani [9] explained that pesticide, temperature, and rainfall were among the most effective predictors for major crops in Saudi Arabia like wheat and sorghum, achieved 0.96 of R^2 and 0.0449 of root mean square error (RMSE) using a multilayer perceptron (MLP) model.

Another recent research by Islam et al. [14] used XGBoost for yield prediction like wheat, maize, sorghum, potatoes and rice in the Saudi Arabia and obtained an R^2 score of 0.97 and RSME of 15803.15. Their research focused on the role of data in promoting climate-resilient agriculture in the region.

Al-Gaadi et al. [25] developed a precision agriculture framework to predict potato yield in Eastern Saudi Arabia using multispectral imagery (Landsat-8, Sentinel-2). Vegetation indices (NDVI, SAVI, CNDVI, CSAVI) were correlated with field yield data from 30-ha plots. Linear regression models showed Sentinel-2 performed best ($R^2 = 0.65$, RMSE = 4.96%), with optimal prediction 60–70 days post-planting. The study highlights the effectiveness of high-resolution satellite monitoring for yield estimation in arid regions.

Al-Gaadi et al. [26] assessed the water footprint (WF) and productivity of carrots and onions in arid Saudi conditions using Sentinel-2, Landsat-8, and the SSEB model. Spectral indices (NDVI, SAVI, RDVI, EVI) and LST were used to predict yield and evapotranspiration. Linear regression with NIR band achieved $R^2 = 0.77$ (carrots) and 0.68 (onions). Estimated WF was 312 m^3/t (carrots) and 230 m^3/t (onions), confirming the value of satellite-based monitoring for yield and water use efficiency.

Li et al. [27] performed a national-scale analysis of Saudi agriculture (1990–2021) using a hybrid ML framework combining DBSCAN, CNNs, and spectral clustering to map 28,000+ center-pivot fields via Landsat NDVI composites. Delineation accuracy was high ($R^2 > 0.97$; producer's: 83.7–94.8%, user's: 90.2–97.9%). Results showed expansion (2010–2015) and post-2016 contraction tied to water policy. The study offers high-resolution field data for future yield, WF, and crop type studies in arid regions.

Ahmed [28] proposed a maize yield prediction framework for Saudi Arabia using a modified MLP optimized by Spider Monkey Optimization (SMO). Trained on FAO and World Bank data (rainfall, temperature, yield), the MLP-SMO model outperformed LASSO, XGBoost, LightGBM, RF, SVM, GRNN, and LSTM, achieving $R^2 = 0.98$ and RMSE = 0.11 Mg/Ha. Statistical tests (MAE, MBE, Wilcoxon) confirmed its robustness. The study highlights the potential of hybrid neural-metaheuristic models for accurate yield forecasting in arid regions.

Jabbari et al. [29] studied IoT adoption for crop monitoring and yield prediction among 550 farmers in Jizan, Saudi Arabia. Statistical analysis showed a strong correlation between awareness and perceived benefits ($r = 0.835$, $p < 0.001$). Perceived benefits explained 21% of adoption variance ($R^2 = 0.210$), while access to information and government support were also significant predictors ($R^2 = 0.191$). The study highlights the need for targeted training and institutional backing to boost IoT adoption in precision agriculture.

A summary of the approaches for crop yield prediction in Saudi Arabia is given in table I.

TABLE I
EXISTING WORKS OF CROP YIELD PREDICTION IN SAUDI ARABIA

Ref, Year	Crop	Model	R^2	Data Source
[14], 2024	Maize, Potatoes, Sorghum, Soybean, Wheat	Bagging RF XGBoost	$\geq$0.91	FAO + World Bank Open Data
[12], 2023	Dates Maize Potatoe Wheat	Neural Network Neural Network	0.974 0.792 0.499 0.930	FAO + Thier global data
[27], 2023	Wheat, Barley, Alfalfa, Fruits, Vegetables	DBSCAN, CNN (AlexNet), Spectral Clustering	0.97	GEE, Manual digitization
[28], 2023	Maize	MLP + Spider Monkey Optimization (SMO)	0.98	FAO, World Data Bank
[29], 2023	Various	Regression (correlation + MLR)	0.698	Questionnaire
[9], 2022	Potatoes Rice Sorghum Wheat	MLP MLP MLP MLP	0.99 0.90 0.99 0.99	Kaggle
[26], 2022	Carrots	Linear Regression	0.77	USGS, WorldClim.org, Tawdeehiya Farms
[25], 2016	Potatoes	Linear Regression	0.65	USGS, INMA Co.

Despite advances in ML/DL for crop yield prediction, significant gaps exist in the literature, particularly in terms of regional adaptability to arid climates, full ensemble model evaluation, and holistic data fusion methodologies. Existing models frequently lack calibration for regions such as Saudi

Arabia, and therefore struggle with generalizable, scalable architecture. While specific ensemble strategies have been verified, few studies have systematically compared a wide range of ensemble methods across many crops and regions, a gap that this work fills by evaluating seven algorithms for Saudi crops. Furthermore, many implementations miss critical data fusion, particularly the combined influence of environmental, temporal, remote sensing, and management variables, whereas this study distinguishes the varied importance of components such as NDVI versus precipitation in the Saudi context. Also, we claim that the suggested approach has great generalizability, allowing for unlimited use across many situations, provided the necessary data is available.

III. Methodology

Sustainable food production in arid regions like Saudi Arabia faces significant challenges due to extreme climate, limited water resources, and constrained arable land. Crop yield prediction serves as a crucial tool to improve agricultural planning, optimize resources use, and support national food security. While ML and DL methods have shown promise in this domain, most prior studies rely on global datasets and focus primarily on few crops. These approaches often fail to generalize to local conditions and underrepresent crops vital to the Saudi agricultural economy, such as fruits and dates.

This research proposes a framework based on ensemble machine learning tailored to the agro-climatic context of Saudi Arabia. By leveraging multi-source datasets and a range of ensemble models including Random Forest, Bagging Decision Trees, Gradient Boosting, Stochastic Gradient Boosting, XGBoost, AdaBoost, and CatBoost, this study aims to deliver accurate and scalable yield predictions.

As illustrated in figure 1, the methodology is structured as follows: data collection and preprocessing, model training using historical and environmental variables (e.g., temperature, precipitation, NDVI, pesticide usage), performance evaluation using standard regression metrics (R^2, MAE, RMSE, MAPE), and different optimization experiments. This structured approach ensures robust, context-aware model development that supports smart agriculture in arid environments.

A. Data collection

To develop an effective crop yield prediction framework tailored to Saudi Arabia, a diverse set of datasets was compiled from reputable global platforms. The features were selected to reflect environmental, agronomic, and remote sensing variables that have been shown to impact crop productivity in prior studies.

- Crop type and yield: Crop-specific production and yield data were obtained from the the Food and Agriculture Organization (FAO) [30]. The data included total yield values (in kg per hectare) for several crops cultivated in Saudi Arabia like wheat, maize, tomatoes, potatoes, barley, sorghum, and dates.
- Pesticide use: Annual pesticide usage was collected from FAO databases. This variable represents the total quantity of pesticides applied in agricultural operations per year.
- Weather data: Meteorological data, including average annual temperature, total precipitation, and average humidity, were extracted from the World Bank Climate Data API [31]. The data were aggregated on a yearly basis for the studied country.
- Vegetation indices: The Normalized Difference Vegetation Index (NDVI) is a widely used spectral index that quantifies vegetation health and density by measuring the difference between near-infrared (NIR) and red (RED) reflectance captured by satellite sensors [1]. NDVI is calculated using the formula:

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$$

This formula exploits the fact that healthy vegetation reflects more NIR light and absorbs more red light, while sparse or unhealthy vegetation shows the opposite pattern. NDVI values range from -1 to $+1$, where higher positive values typically indicate dense, healthy vegetation, and values near zero or negative indicate barren areas, water bodies, or built-up land.
In this study, NDVI images were extracted and processed using the Google Earth Engine (GEE) platform, which provides cloud-based access to satellite imagery and enables large-scale geospatial analysis. Landsat 5, 7, and 8 Surface Reflectance (SR) collections were utilized to ensure temporal continuity from 1990 to 2022. The satellite images were first filtered for cloud-free conditions and spatially clipped to specific agricultural regions in Saudi Arabia. For each image, NDVI was computed using GEE's built-in `normalizedDifference()` function, which efficiently implements the NDVI formula.

A summary of the data sources is given in Table II.

TABLE II
SUMMARY OF DATA SOURCES USED IN THE STUDY

Source	Type of Data	Time Range
FAO	Crop yield, pesticide use	1990–2022
GEE	Vegetation indices (NDVI)	1990–2023
World Bank Climate Data	Temperature, humidity, precipitation	1990–2022

B. Data pre-processing

Following the collection of raw datasets from various sources, basic pre-processing steps were applied to prepare the data for ML modeling. These steps were essential to ensure the consistency, completeness, and relevance of the input features.

- Vegetation indices computation:
To obtain a comprehensive overview of the crop condition, NDVI is used along with VCI (Vegetation Condition Index) and WDRVI (Wide Dynamic Range Vegetation Index). The three indices give complementary vegetation

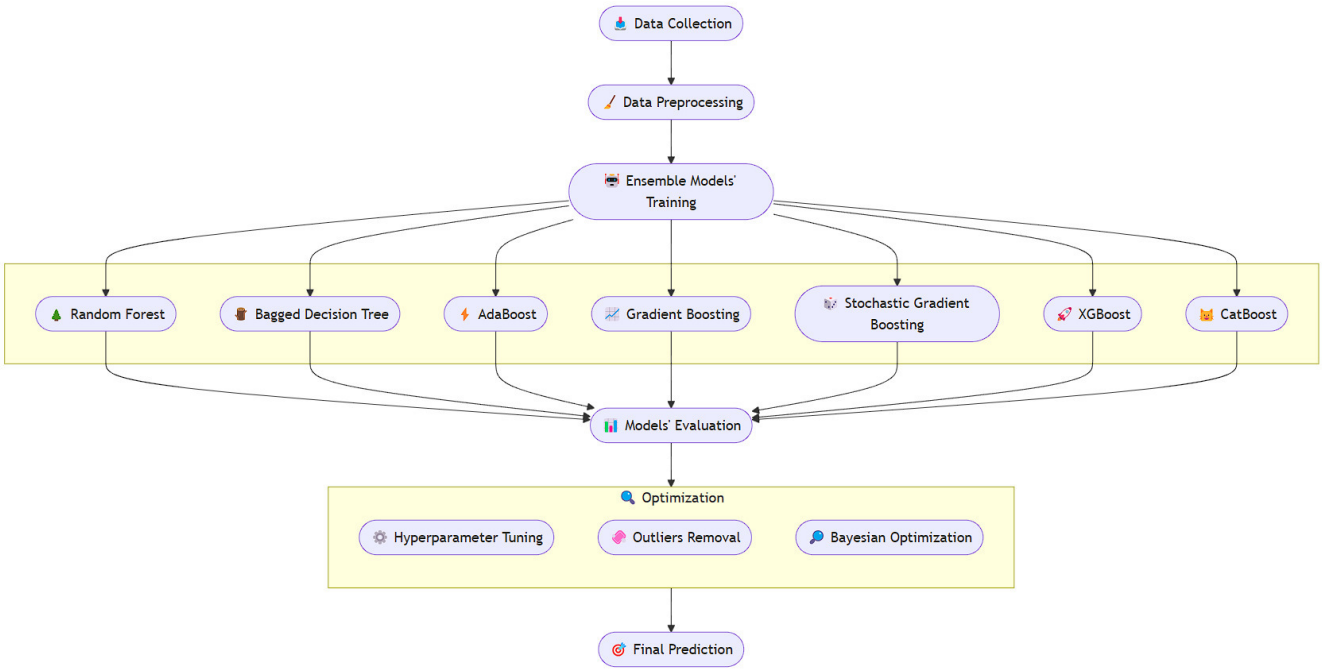

Fig. 1. Proposed framework

information, which is notably significant in agricultural surveillance and crop yield forecasting.

- VCI was calculated using the standard formula [4]:

$$VCI = \frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \times 100$$

- WDRVI was derived from red and near-infrared reflectance bands using:

$$WDRVI = \frac{(0.1 \cdot NIR) - RED}{(0.1 \cdot NIR) + RED}$$

where NIR and RED are the near-infrared and red reflectance values from MODIS [4].

- Data aggregation: All variables were aggregated at the annual level to align crop yield values with climate, pesticide, and remote sensing data.
- Missing data handling: Records with missing values were dropped or imputed using mean imputation where appropriate. This was particularly relevant in the early years with sparse pesticide data.
- Categorical encoding: The crop name (such as Wheat, Sorghum, Dates) was converted into numeric labels using label encoding to allow its use as a categorical feature in selected models. This is a critical step since machine learning models need numerical data. For presentation and evaluation, the original harvest names are kept in distinct variables or extracted using inverse_transform.

All datasets were harmonized to a common temporal resolution (annual), cleaned to remove missing values, and merged into a unified data frame containing 1117 records across all crop types and years. Descriptive data statistics are illustrated in figure 2.

C. Ensemble models

This research proposes a multi-stage ML pipeline to predict crop yields in various types of crop cultivated in Saudi Arabia. The methodology is centered on ensemble techniques, given that ensemble machine learning takes advantage of the complementary characteristics of many models to improve predictive accuracy, minimize variation and bias, and achieve improved generalization in complicated tasks. We implemented seven from the famous bagging and boosting models and proposed different optimizations based on regularization and hyperparameter tuning applied to the seven models. The analysis involved the evaluation of each model across all crops in the dataset, then the selection of the top-performing crops on which the optimization techniques were tested, and finally the in-depth training, tuning, and evaluation of all models per selected crop. The seven applied ensemble learning models fall in the bagging and boosting ensemble categories:

- Bagging models: Bagging models increase the accuracy of predictions by training several base learners simultaneously on random portions of the data.
  - Random Forest (RF) is an ensemble method that constructs multiple decision trees on bootstrapped samples of the data and averages their outputs. For regression, the prediction is given by:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} T_i(x)$$

	Year	Yield_Value	Temperature	Precipitation	Humidity	NDVI_Mean
count	1117.00	1117.00	1117.00	1117.00	1117.00	1117.00
mean	2006.93	12690.34	25.59	60.67	31.16	0.05
std	9.79	11713.60	0.41	5.93	0.45	0.01
min	1990.00	96.80	24.55	52.69	30.40	0.04
25%	1998.00	3974.50	25.29	56.09	30.81	0.04
50%	2007.00	10000.00	25.60	59.47	31.09	0.04
75%	2016.00	18946.80	25.88	64.34	31.41	0.05
max	2022.00	85800.50	26.21	75.59	32.04	0.06

	VCI	WDRVI_Mean	Pesticides_Value
count	1117.00	1117.00	1117.00
mean	46.48	0.03	4653.51
std	32.52	0.00	2717.36
min	0.00	0.02	994.00
25%	15.99	0.02	2683.37
50%	36.04	0.03	4216.71
75%	81.06	0.03	7255.65
max	100.00	0.03	10495.54

Fig. 2. Collected Data Statistics

where $T_i(x)$ is the prediction from the i-th DT and K is the total number of trees [14].

- Bagged Decision Tree (Bagged DT) is similar to RF but without feature randomness; trees are trained on bootstrapped datasets only.

- Boosting models: Boosting models construct ensembles in stages, with each model focusing on rectifying the faults of the previous ones.
  - Gradient Boosting (GB) iteratively minimizes a loss function via gradient descent:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

  - Stochastic Gradient Boosting (Stochastic GB) is a variant of GB that introduces randomness by sampling the training set at each iteration [20].
  - eXtreme Gradient Boosting (XGBoost) is an optimized version of GB with regularization:

$$\text{Obj} = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t)$$

    where $\Omega(f_t)$ is a complexity penalty for tree f_t [20].
  - AdaBoost sequentially trains weak learners and combines them with weights:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

    where h_m is the m-th weak learner and α_m is its weight.
  - CatBoost uses ordered boosting and handles categorical features natively to reduce prediction shift and overfitting [20].

This ensemble learning-based framework proposed for crop yield prediction was inspired by the work of Zhou [20], which details the theoretical underpinnings and practical successes of such models in various domains.

All seven models were trained and evaluated in their baseline definition (called baseline models in the sequel), then three optimizations were proposed for each model resulting in 21 other models named respectively OR, Hypertuned, and BayesSearchCV for models optimized using Outliers Removal, Hyperparameters tuning, and Bayesian optimization.

D. Model's optimization

To enhance the accuracy and generalizability of the ensemble models, a structured optimization pipeline was implemented. This process consisted of three sequential steps: hyperparameter tuning, outlier removal, and Bayesian optimization. Each step was selected to address specific challenges in predictive modeling and improve the models' robustness.

1) Hyperparameter tuning: The optimization phase began with hyperparameter tuning using GridSearchCV, a technique that systematically explores combinations of predefined hyperparameter values. This approach enables the selection of optimal configurations for each model, such as the number of estimators, tree depth, and learning rate, by evaluating model performance across cross-validated folds.

Hyperparameter tuning plays a crucial role in reducing underfitting or overfitting, and ensures that models are well-calibrated for the underlying data distribution. Its effectiveness is well-documented in the ML literature [22].

In this study, a distinct set of parameter grids was defined for each ensemble model to account for their structural and functional differences. The search space proposed for each model, the best parameters obtained, and the crop for which the best values are achieved are summarized in table III.

2) Outliers' removal: As an *indirect* optimization method, we relied on the outliers' removal method to clean the data by removing anomalies, which identifies and excludes data points that differ considerably from the dataset's anticipated trends or distribution. This may lead to better generalization and lower error metrics (e.g., RMSE and MAE).

TABLE III
SELECTED HYPERPARAMETERS FOR TUNING THE DEVELOPED ENSEMBLE MODELS

Model	Search Space	Best Parameters	Crop
Random Forset	n_estimators: [50, 100, 200, 300]	50	Pulses, Total
	max_depth: [None, 10, 20]	None	
	min_samples_split: [2, 5, 10]	2	
	min_samples_leaf: [1, 3]	1	
	max_features: ['sqrt', 'log2', 0.8]	0.8	
AdaBoost	n_estimators: [50, 100, 200, 300, 500]	300	Pulses, Total
	learning_rate: [0.001, 0.01, 0.1, 0.5]	0.001	
	loss: ['linear', 'square', 'exponential']	exponential	
XGBoost	n_estimators: [100, 200, 300]	200	Pulses, Total
	learning_rate: [0.02, 0.05, 0.1, 0.2]	0.2	
	subsample: [0.8, 0.9, 1.0]	0.8	
	colsample_bytree: [0.8, 0.9, 1.0]	0.8	
	max_depth: [3, 4, 5]	3	
	reg_alpha: [0, 0.001, 0.01, 0.1]	0.1	
	reg_lambda: [1, 0.1, 0.5, 1.0]	0.1	
CatBoost	iterations: [100, 200, 300]	300	Pulses, Total
	learning_rate: [0.01, 0.05, 0.1, 0.2]	0.05	
	depth: [3, 4, 5]	4	
	l2_leaf_reg: [1.0, 3.0, 5.0]	5.0	
Bagged Decision Tree	n_estimators: [10, 50, 100, 200]	10	Pulses, Total
	max_samples: [0.5, 0.6, 0.8, 1.0]	1.0	
	max_features: [0.5, 0.6, 0.8, 1.0]	1.0	
	estimator__max_depth: [None, 5, 10, 15]	None	
Gradient Boosting	n_estimators: [100, 200,300,500]	100	Fruit Primary
	learning_rate: [0.05, 0.1, 0.2]	0.1	
	subsample: [0.6, 0.8, 1.0]	0.8	
	max_depth: [3, 4, 5]	3	
Stochastic Gradient Boosting	n_estimators: [100, 200, 300, 400]	400	Pulses, Total
	learning_rate: [0.05, 0.1, 0.2]	0.05	
	subsample: [0.7, 0.8, 0.9]	0.8	
	max_depth: [3, 4, 5]	3	
	min_samples_split: [2, 5]	2	
	min_samples_leaf: [1,3]	3	

Once the models were tuned, Isolation Forest was applied to remove outliers from the data. This unsupervised anomaly detection technique isolates unusual observations by randomly partitioning the feature space. Its strength lies in its scalability and independence from the data distribution assumptions.

To maintain consistency, the cleaned dataset was passed into retrained versions of the same ensemble models, each with fixed hyperparameters tailored for robust generalization. For instance, RF and Bagged DT were retrained using 100 estimators, while AdaBoost used 100 weak learners. GB and SGB were configured with 200 and 100 estimators respectively, with subsampling enabled to enhance robustness. XGBoost was applied with additional regularization (`reg_alpha` = 0.1, `reg_lambda` = 1.0) and feature subsampling (`colsample_bytree` = 0.8), aiming to mitigate overfitting on reduced data. CatBoost was configured with 200 iterations, a moderate learning rate (0.05), and a regularization term (`l2_leaf_reg` = 3.0), while disabling verbose output for computational efficiency.

All models used a consistent random seed to ensure reproducibility across experimental runs. Outliers' removal helped to mitigate the impact of noise and extreme values, which can distort model learning and inflate error metrics. This is particularly beneficial in agricultural datasets where variability is high [23].

3) Bayesian optimization: In the final step, Bayesian Optimization was conducted using `BayesSearchCV`. Unlike grid search, which exhaustively tests all combinations, Bayesian optimization models the objective function and iteratively selects hyperparameters that are likely to yield the best results. This method improves the search efficiency, particularly in high-dimensional spaces, and has been shown to outperform traditional tuning methods in many ML applications [24].

For this study, specific search spaces for each model were designed to reflect their structural characteristics and known sensitivities. The parameters used in these spaces are defined in table IV. In the same table, we report the best values obtained and the crop for which the model performs better.

These customized search spaces allowed for targeted exploration of the hyperparameter landscape while maintaining computational efficiency. All models were optimized with a fixed random seed for reproducibility, and verbosity was disabled for resource-controlled environments.

IV. RESULTS AND DISCUSSION

Rigorous assessment of the developed models is critical for verifying the efficacy of models, ensuring repeatability, and quantifying predicted reliability, which helps to enable accurate choices and trustworthy real-world deployment.

Our objective was to develop and evaluate seven ensemble machine learning models in their base configuration in addition

TABLE IV
SELECTED HYPERPARAMETERS FOR BAYESIAN OPTIMIZATION OF THE DEVELOPED ENSEMBLE MODELS

Model	Search Space	Best Parameters	Crop
Random Forset	n_estimators: (100, 1000) max_depth: (3, 25) max_features: (0.1, 1.0)	100 3 0.9274781325161314	Pulses, Total
AdaBoost	n_estimators: (50, 500) learning_rate' (0.005, 2.0, loguniform) loss: ['linear', 'square', 'exponential']	53 0.018596594453385986) exponential	Pulses, Total
XGBoost	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, loguniform) max_depth: (3, 15) subsample: (0.5, 1.0) colsample_bytree: (0.5, 1.0)	384 0.1426991843302091 14 0.8350739741344673 0.705051979426657	Pulses, Total
CatBoost	iterations: (100, 1000) learning_rate: (0.005, 0.5, loguniform) depth: (3, 15) l2_leaf_reg: (0.1, 20.0, loguniform) border_count: (32, 255)	898 0.035612784787118226 3 1.2563877010366442 59	Other fruits, n.e.c.
Bagged Decision Tree	n_estimators: (10, 200) max_samples: (0.5, 1.0) max_features: (0.5, 1.0) estimator__max_depth: (3, 20)	10 1.0 1.0 3	Fruit Primary
Gradient Boosting	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, log-uniform) max_depth: (3, 15)	247 0.14690143571379652 14	Pulses, Total
Stochastic Gradient Boosting	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, log-uniform) max_depth: (3, 15) subsample: (0.5, 1.0) min_samples_split: (2, 10) min_samples_leaf: (1, 10)	269 0.03878243104032394 14 0.7268326719031495 5 2	Pulses, Total

to the application of three techniques to study the models' performance enhancement. This resulted in the evaluation of 28 models resulting from the seven ensemble models discussed earlier multiplied by the four possible configurations. For this, we chose to assess these models on selected crops instead of studying all the crops obtained in the collected data. This choice is made on the basis of training the seven baseline models on each crop in the dataset and selecting the eight crops with best results according to the following metrics:

- R^2: Measures the proportion of variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating better model fit:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Root Mean Squared Error (RMSE): is a fundamental evaluation metric widely used in regression problems, including crop yield prediction. It quantifies the square root of the average of the squared differences between predicted values $\hat{y}_i$ and actual values y_i, and is formally defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n denotes the number of observations. The squaring of errors penalizes larger deviations more than smaller ones, making MSE particularly sensitive to outliers. This sensitivity allows it to highlight substantial prediction errors that may impact agricultural decision making.

- Mean Absolute Error (MAE): Represents the average magnitude of the prediction errors, without considering their direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Absolute Percentage Error (MAPE): is a regression metric that measures the average absolute percentage difference between predicted and actual values, making it useful for determining the accuracy of the model

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

The crops selected for further investigation and the metrics obtained for the best model for each crop are summarized in table V.

A. Baseline models' Performance

As mentioned in the previous section, to establish a benchmark for model performance, all ensemble models were initially evaluated using default (untuned) configurations. The evaluation was carried out across eight diverse crop categories using standard regression metrics: MAE, RMSE, R^2, and MAPE (table V). We aim here to provide more insights about the training of the seven ensemble models (with default configurations) on the selected crops.

TABLE V
BEST MODEL AND ITS EVALUATION METRICS FOR EACH CROP

Crop	Best Model	MAE	RMSE	R^2	MAPE
Cereals, primary	AdaBoost (Baseline)	0.015	0.019	0.829	0.042
Dates	Gradient Boosting (Baseline)	0.022	0.034	0.861	0.299
Fruit Primary	Stochastic Gradient Boosting (Baseline)	0.026	0.035	0.918	0.890
Other fruits, n.e.c.	Stochastic Gradient Boosting (Baseline)	0.075	0.103	0.923	0.252
Pulses, Total	AdaBoost (Baseline)	0.003	0.005	0.976	0.007
Sorghum	Stochastic Gradient Boosting (Baseline)	0.008	0.013	0.894	0.016
Wheat	Stochastic Gradient Boosting (Baseline)	0.014	0.019	0.864	0.051
Pumpkins, squash and gourds	Random Forest (Baseline)	0.088	0.119	0.880	0.534

To visualize the overall model performance across crops, figure 3 presents a heatmap of R^2 scores, capturing the comparative behavior of all base models.

The heatmap confirms that CatBoost and Stochastic GB seem to be strong performers overall, consistently showing high R^2 scores. AdaBoost and Gradient Boosting show very good performance across most crops. Random Forest generally performs well achieving 97% for Pulses crop. On the other hand, Bagged DT and XGBoost appear to have lower R^2 scores, particularly for Dates and Other fruits, which supposes that they might not be suitable for these crop types compared to the other models.

From the lens of crops, Pulses, Fruit, and Sorghum seem to be relatively well-predicted across most models, with many R^2 scores greater than 0.80 and a score ranging between 0.91 and 0.98 for Pulses crop. Whereas, Dates and Other fruits tend to have lower R^2 scores for several models, indicating that predicting the target variable for these crops might be more challenging.

Learning curves obtained by testing the models' effectiveness on progressively bigger subsets of the input data through cross-validation were studied to identify probable variance and bias concerns. To understand how the seven models learn and generalize from data, we visualize the learning curves for "Pulses, Total" and "Dates". For Pulses crop, figure 4 shows that the models are capable of extremely quick learning. Despite limited training instances (e.g., 5-10), most models get strong R^2 values (typically greater than 0.80) on both training and cross-validated data. The curves increase rapidly and converge soon. However, figure 5 shows that for Dates, the models demonstrate slower and less efficient learning. Low training instances (e.g., up to 15-20) show a significant disparity between training R^2 (typically 1.0) and CV R^2 (near 0 or even negative), indicating early overfitting.

B. Performance of the optimized Models

For direct and indirect optimization of the developed models across the selected crops, we performed the following tasks as explained in the methodology section.

- Hyperparameter tuning
- Outliers' removal
- Bayesian optimization

The R^2 scores for all models are summarized in table VI. Accoding to the table, the proposed optimizations have significantly improved the models' performance across the studied crops, particularly for formerly underperforming predictions. For Dates crop for example, the XGBoost model showed an increase in R^2 score by $\tilde{4}7\%$ after hyperparameter tuning. Also, Bayesian optimization enhances dramatically the prediction of Dates by Bagged Decision Tree (R^2 scores passes from 0.67 to 0.92). It is indeed clear from the table that almost all the optimizated models perform better than the default models.

We also highlighted in gray the cell showing the best R^2 score for each crop for model selection. For example, the best predictor for Fruits is the hypertuned Stochastic GB model with a score R^2 of 0.98.

The capacity to accomplish such consistent generalization over a broader range of crop varieties demonstrates the success of our optimization method and establishes these models as dependable tools for the forecasting of crop yields.

V. CONCLUSION

In this research, we developed and evaluated seven ensemble-based machine learning models to predict eight crop yields in Saudi Arabia. In order to enhance the baseline models' performance, three optimization methods were developed: direct optimization via hyperparameter tuning and Bayesian optimization and indirect optimization through outliers' removal.

The results demonstrated that Stochastic Gradient Boosting and XGBoost are the best performers, each model achieving the highest R^2 values for three different crops. Then, CatBoost and AdaBoost were champions in predicting Other fruits and Cereals, resp. However, bagging models (RF and Bagged DT), and the GB model did not outperform the developed models in any of the selected crops.

It should be noted that the optimization strategies had varied effects. Hyperparameter tuning showed marginal gains, while Bayesian optimization and outlier removal (via Isolation Forest) led to noticeable performance improvements. However, the effectiveness of each strategy was dependent on the crop.

In summary, the proposed framework provided a scalable and interpretable approach to smart agriculture in Saudi Arabia. Its crop-level insights support data-driven decision-making for farmers and policymakers.

For future investigations, it is promising to expand the dataset to include soil characteristics, crop management practices, and economic indicators to improve model generalization. Also, it is propitious to explore deep learning architec-

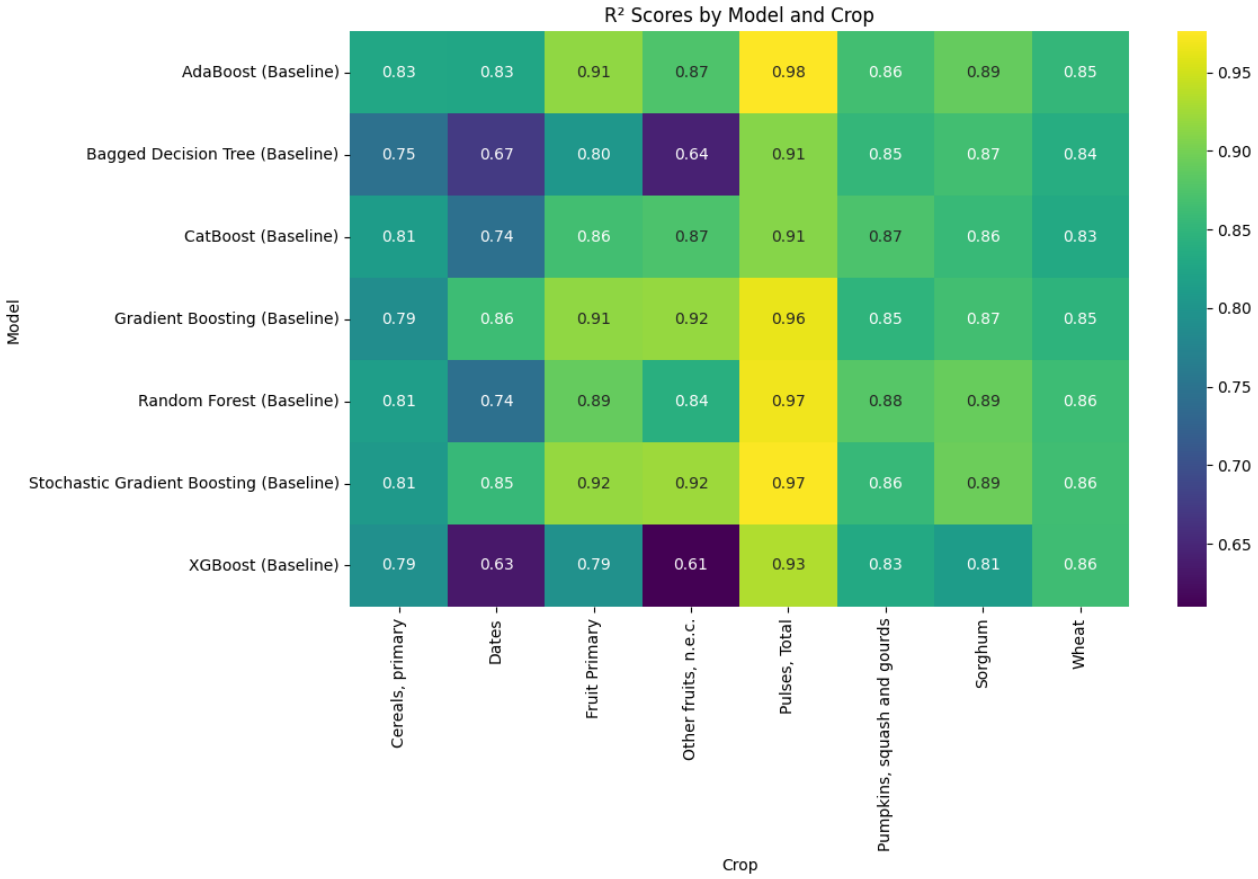

Fig. 3. Heatmap of R^2 values for baseline models across crops

tures such as LSTM or CNN to capture temporal dynamics more effectively where larger amount of data is available. Finally, exploring the performance of deep ensemble learning on the studied crops appears promising, given the encouraging results reported in [33].

REFERENCES

[1] M. Ashfaq, I. Khan, A. Alzahrani, M. U. Tariq, H. Khan, and A. Ghani, *Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion*, IEEE Access, vol. 12, pp. 40947–40961, 2024. doi:10.1109/ACCESS.2024.3376735.

[2] A. D. Yewlea, L. Mirzayeva, and O. Karakuş, *Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction*, Preprint submitted to Elsevier, Feb. 2025.

[3] A. Joshi, B. Pradhan, S. Chakraborty, R. Varatharajoo, S. Gite, and A. Alamri, *Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data*, Remote Sensing, vol. 16, no. 24, 4804, 2024. https://doi.org/10.3390/rs16244804.

[4] Q. M. Ilyas, M. Ahmad, and A. Mehmood, *Automated Estimation of Crop Yield Using Artificial Intelligence and Remote Sensing Technologies*, Bioengineering, vol. 10, no. 2, article 125, pp. 1–24, 2023. doi:10.3390/bioengineering10020125.

[5] P. Brandt, F. Beyer, P. Borrmann, M. Möller, and H. Gerighausen, *Ensemble learning-based crop yield estimation: a scalable approach for supporting agricultural statistics*, GIScience & Remote Sensing, vol. 61, no. 1, pp. 2367808, 2024. DOI: 10.1080/15481603.2024.2367808.

[6] Y. Wang, Q. Zhang, F. Yu, N. Zhang, X. Zhang, Y. Li, M. Wang, and J. Zhang, *Progress in Research on Deep Learning-Based Crop Yield Prediction*, Agronomy, vol. 14, no. 10, article 2264, pp. 1–26, 2024. doi:10.3390/agronomy14102264.

[7] J. Ansarifar, L. Wang, and S. V. Archontoulis, *An Interaction Regression Model for Crop Yield Prediction*, Scientific Reports, vol. 11, article 17754, 2021. doi:10.1038/s41598-021-97221-7.

[8] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, *A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction*, IEEE Access, vol. 9, pp. 63406–63439, 2021.

[9] M. H. Al-Adhaileh and T. H. H. Aldhyani, *Artificial Intelligence Framework for Modeling and Predicting Crop Yield to Enhance Food Security in Saudi Arabia*, PeerJ Computer Science, vol. 8, article e1104, 2022. doi:10.7717/peerj-cs.1104.

[10] F. M. Talaat, *Crop Yield Prediction Algorithm (CYPA) in Precision Agriculture Based on Internet of Things (IoT) Techniques and Climate Changes*, Neural Computing and Applications, vol. 35, pp. 17281–17292, 2023. doi:10.1007/s00521-023-08619-5.

[11] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, *Crop Prediction Using Machine Learning*, Journal of Physics: Conference Series, vol. 2161, 012033, 2022. doi:10.1088/1742-6596/2161/1/012033.

[12] H. F. Assous, H. AL-Najjar, N. Al-Rousan, and D. AL-Najjar, *Developing a Sustainable Machine Learning Model to Predict Crop Yield in the Gulf Countries*, Sustainability, vol. 15, no. 12, article 9392, 2023. doi:10.3390/su15129392.

[13] K. Palanivel and C. Surianarayanan, *An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques*, International Journal of Computer Engineering and Technology (IJCET), vol. 10, no. 3, pp. 110–118, 2019.

[14] M. M. Islam, M. Alharthi, R. S. Alkadi, R. Islam, and A. K. M. Masum, *Crop Yield Prediction through Machine Learning: A Path Towards Sustainable Agriculture and Climate Resilience in Saudi Ara-*

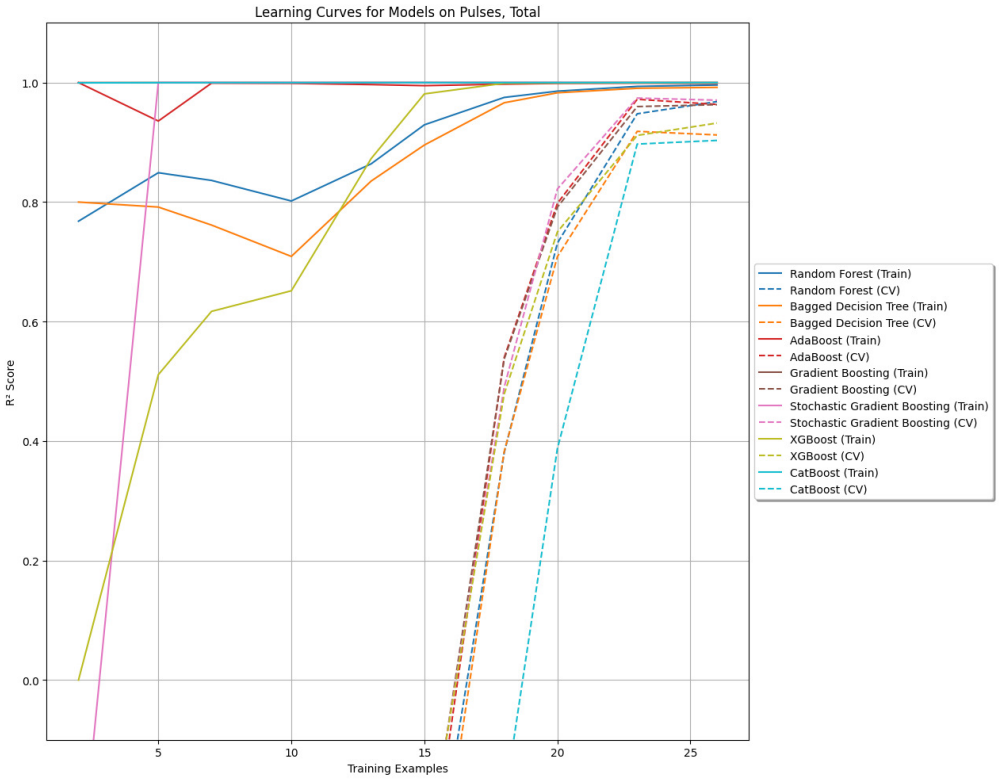

Fig. 4. Learning Curves for models on the crop: ”Pulses, Total”

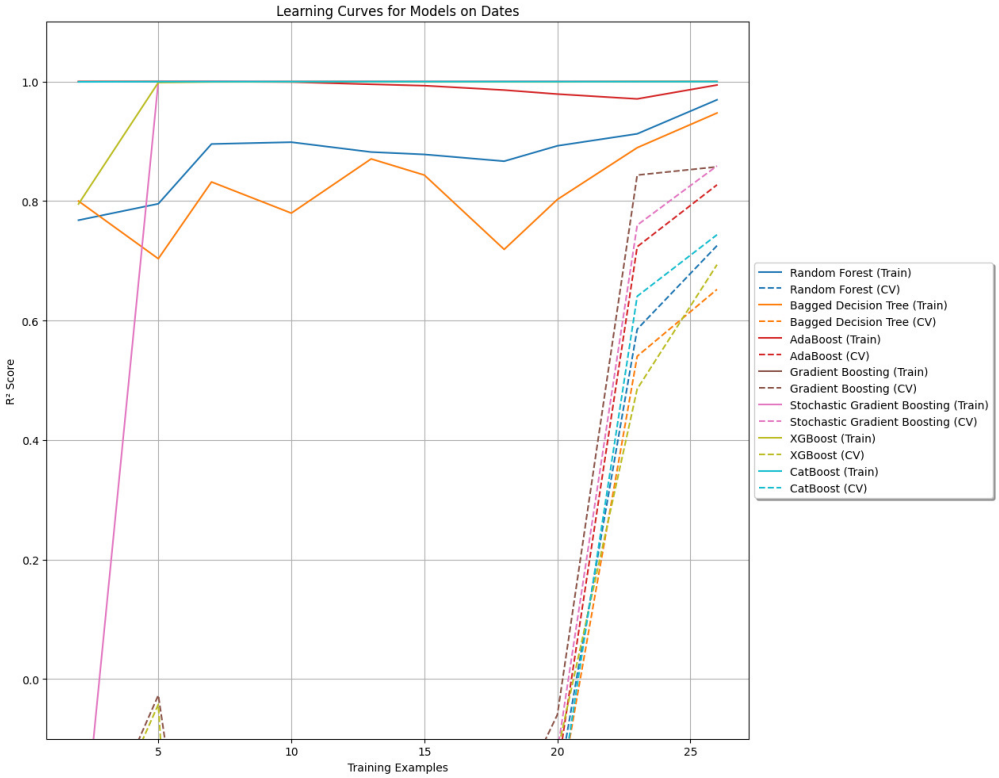

Fig. 5. Learning Curves for models on the crop: ”Dates”

TABLE VI
R^2 SCORE FOR EACH MODEL PER CROP

Model	Cereals, primary	Dates	Fruit Primary	Other fruits, n.e.c.	Pulses, Total	Sorghum	Wheat	Pumpkins, squash and gourds
AdaBoost (Baseline)	0.829	0.826	0.915	0.873	0.976	0.889	0.852	0.865
AdaBoost (Hypertuned)	0.775	0.951	0.961	0.943	0.983	0.807	0.816	0.893
AdaBoost (OR)	0.908	0.919	0.899	0.825	0.975	0.767	0.889	0.915
AdaBoost (BayesSearchCV)	0.827	0.947	0.962	0.947	0.982	0.828	0.827	0.893
Bagged Decision Tree (Baseline)	0.745	0.672	0.803	0.644	0.909	0.865	0.838	0.852
Bagged Decision Tree (Hypertuned)	0.804	0.915	0.940	0.890	0.951	0.883	0.828	0.932
Bagged Decision Tree (OR)	0.897	0.885	0.900	0.891	0.971	0.906	0.885	0.946
Bagged Decision Tree (BayesSearchCV)	0.823	0.926	0.931	0.866	0.928	0.861	0.853	0.930
CatBoost (Baseline)	0.811	0.744	0.865	0.873	0.910	0.856	0.831	0.872
CatBoost (Hypertuned)	0.705	0.896	0.923	0.963	0.971	0.863	0.906	0.911
CatBoost (OR)	0.846	0.735	0.825	0.885	0.949	0.905	0.841	0.959
CatBoost (BayesSearchCV)	0.692	0.866	0.921	0.963	0.963	0.849	0.914	0.931
Gradient Boosting (Baseline)	0.789	0.861	0.915	0.918	0.963	0.865	0.847	0.848
Gradient Boosting (Hypertuned)	0.767	0.956	0.981	0.930	0.980	0.830	0.889	0.881
Gradient Boosting (OR)	0.852	0.895	0.956	0.953	0.981	0.829	0.865	0.919
Gradient Boosting (BayesSearchCV)	0.747	0.952	0.967	0.834	0.983	0.847	0.621	0.892
Random Forest (Baseline)	0.813	0.743	0.889	0.841	0.968	0.891	0.862	0.880
Random Forest (Hypertuned)	0.773	0.906	0.941	0.891	0.953	0.846	0.869	0.948
Random Forest (OR)	0.904	0.866	0.879	0.919	0.974	0.913	0.880	0.953
Random Forest (BayesSearchCV)	0.829	0.927	0.946	0.892	0.974	0.873	0.868	0.930
Stochastic GB (Hypertuned)	0.694	0.958	0.983	0.936	0.993	0.856	0.907	0.887
Stochastic GB (Baseline)	0.807	0.855	0.918	0.923	0.974	0.894	0.864	0.855
Stochastic GB (OR)	0.848	0.912	0.934	0.941	0.967	0.850	0.861	0.911
Stochastic GB (BayesSearchCV)	0.799	0.960	0.977	0.949	0.985	0.821	0.898	0.804
XGBoost (Baseline)	0.793	0.634	0.793	0.610	0.933	0.811	0.862	0.831
XGBoost (Hypertuned)	0.784	0.933	0.860	0.833	0.961	0.845	0.802	0.911
XGBoost (OR)	0.875	0.886	0.872	0.638	0.959	0.920	0.828	0.980
XGBoost (BayesSearchCV)	0.696	0.814	0.821	0.767	0.959	0.839	0.927	0.896

bia, AIMS Agriculture and Food, vol. 9, no. 4, pp. 980–1003, 2024. doi:10.3934/agrfood.2024053.

[15] A. Morales and F. J. Villalobos, *Using Machine Learning for Crop Yield Prediction in the Past or the Future*, Frontiers in Plant Science, vol. 14, article 1128388, 2023. doi:10.3389/fpls.2023.1128388.

[16] S. V. Joshua, A. S. M. Priyadharson, R. Kannadasan, A. A. Khan, W. Lawanont, F. A. Khan, A. U. Rehman, and M. J. Ali, *Crop Yield Prediction Using Machine Learning Approaches on a Wide Spectrum*, Computers, Materials and Continua, vol. 72, no. 3, pp. 5663–5679, 2022. doi:10.32604/cmc.2022.027178.

[17] C. Trentin, Y. Ampatzidis, C. Lacerda, and L. Shiratsuchi, *Tree Crop Yield Estimation and Prediction Using Remote Sensing and Machine Learning: A Systematic Review*, Smart Agricultural Technology, vol. 9, article 100556, 2024. doi:10.1016/j.atech.2024.100556.

[18] T. van Klompenburg, A. Kassahun, and C. Catal, *Crop yield prediction using machine learning: A systematic literature review*, Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.

[19] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, *Crop Yield Prediction using Machine Learning and Deep Learning Techniques*, Procedia Computer Science, vol. 218, pp. 406–417, 2023.

[20] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series, CRC Press, 2012.

[21] E.-S. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, *Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture*, Potato Research, vol. 68, pp. 759–792, 2025. doi:10.1007/s11540-024-09753-w.

[22] B. Bischl, M. Binder, P. Lang, T. Pfisterer, and A. Richter, *Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 11, pp. 12785–12810, 2023. doi:10.1109/TPAMI.2023.3243326.

[23] G. Pang, C. Shen, L. Cao, and A. van den Hengel, *Deep Learning for Anomaly Detection: A Review*, ACM Computing Surveys, vol. 54, no. 10, Article 200, pp. 1–38, Jan. 2020. doi:10.1145/3439950.

[24] J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 2951–2959, 2012.

[25] K. A. Al-Gaadi, R. Madugundu, E. Tola, and S. El-Hendawy, *Prediction of Potato Crop Yield Using Precision Agriculture Techniques in the Eastern Region of Saudi Arabia*, Journal of the Saudi Society of Agricultural Sciences, vol. 15, no. 2, pp. 82–89, 2016.

[26] K. A. Al-Gaadi, R. Madugundu, E. Tola, S. El-Hendawy, and S. Marey, *Satellite-Based Determination of the Water Footprint of Carrots and Onions Grown in the Arid Climate of Saudi Arabia*, Remote Sensing, vol. 14, no. 23, pp. 1–20, 2022.

[27] T. Li, O. M. L. Valencia, K. Johansen, and M. F. McCabe, *A Retrospective Analysis of National-Scale Agricultural Development in Saudi Arabia from 1990 to 2021*, Remote Sensing, vol. 15, no. 3, pp. 1–21, 2023.

[28] S. Ahmed, *A Software Framework for Predicting the Maize Yield Using Modified Multi-Layer Perceptron*, Sustainability, vol. 15, no. 4, pp. 3017, 2023.

[29] A. Jabbari, A. Humayed, F. A. Reegu, M. Uddin, Y. Gulzar, and M. Majid, *Smart Farming Revolution: Farmer's Perception and Adoption of Smart IoT Technologies for Crop Health Monitoring and Yield Prediction in Jizan, Saudi Arabia*, Sustainability, vol. 15, no. 19, pp. 14541, 2023, doi: 10.3390/su151914541.

[30] https://www.fao.org/faostat/en/#home, last accessed on May 2025.

[31] https://climateknowledgeportal.worldbank.org/, last accessed on May 2025.

[32] G. Ignesti, D. Moroni, M. Martinelli, *Towards the actual deployment of robust, adaptable, and maintainable AI models for sustainable agriculture*, Position Papers of the 19thConference on Computer Science and Intelligence Systems (FedCSIS), 2024, doi: 10.15439/2024F2991.

[33] Z. Sbai, *Deep Learning Models and Their Ensembles for Robust Agricultural Yield Prediction in Saudi Arabia*, Sustainability, vol. 17, no. 13, pp. 1-26, 2025, doi: 10.3390/su17135807.