Hybrid Machine Learning and Metaheuristic Optimization for Hickory Yield Prediction: An Empirical Evaluation of SVR, XGBoost, and RF with AEO and ArchOA

Xu Guo
College of Electronics and Information, Shanghai Dianji University, Shanghai 201306, China
E-mail: guox@sdju.edu.cn
*Corresponding author

Keywords: hickory yield forecasting, RF, SVR, XGBoost, AEO, ArchOA, crop yield, composite schemes, predictive accuracy, execution time

Received: February 11, 2025

Accurate crop yield predictions are vital for sustainable agriculture and resource efficiency, benefiting farmers, agronomists, traders, and policymakers by informing decisions on planting, harvesting, management, trading strategies, and policy. This study explores advanced machine learning methods RF, SVR, and XGBoost for forecasting hickory yields, enhanced by hyperparameter optimization using AEO and ArchOA. Developing and evaluating four predictive schemes, the study employs an 80% training and 20% testing data split for robustness and accuracy. The dataset consists of 52 samples, with variables such as temperature, rainfall, and other related factors. Initial RF, SVR, and XGBoost evaluations are followed by hybrid schemes integrating their strengths for improved accuracy, further refined through systematic hyperparameter optimization. Evaluation metrics include RMSE (Root Mean Squared Error) and R^2 (Coefficient of Determination). In the training phase, XGBoost-AEO had the best performance with an RMSE value of 69.61354 and R^2 of 0.999619. In the testing phase, the XGBoost-AEO model outperformed other models with RMSE of 742.607 and R^2 of 0.959451. Results demonstrate the superior performance of hybrid schemes, especially SVR-AEO and XGBoost-AEO, highlighting the effectiveness of advanced machine learning and optimization techniques in enhancing crop yield predictions and supporting sustainability and food security objectives.

Povzetek: Raziskava združuje strojno učenje in metahevristično optimizacijo za natančno napoved pridelka orehovca, s poudarkom na hibridnih modelih XGBoost-AEO in SVR-AEO za trajnostno kmetijstvo.

1 Introduction

This is a formidable but complex task to be met in pursuit of the sustainable intensification of agriculture and the best use of natural resources. As pointed out in [1][2], crop yield forecasting capability is imperative for the environment and economically sustainable agricultural practice. The accuracy of crop yield forecasts is of immense value to almost all categories of stakeholders involved in the agri-food sector. These stakeholders include farmers, who depend on yield forecasts in making informed decisions on planting and harvesting; agronomists, who use such forecasts to advise on the best practices in managing crops; commodity traders, for whom yield forecasts form the basis of trading strategies and market expectations; and the policymakers, who utilize such data in formulating agricultural policies and food security strategies [3][4]. The interplay between all these factors for ranges of crop yield, being related to climatic conditions, soil health, and agricultural practices, renders the task of developing reliable prediction schemes somewhat sophisticated. This would naturally point to the demand for an integration of various sources of data with higher-order data analytics, which would provide actionable insight to the stakeholder. Improved crop yield forecasting enables the optimization of resource utilization and contributes to attaining general sustainability and food security objectives, given the increasing global demand. Yield forecasting is complex due to the high number of interrelated variables, including soil conditions (e.g., nutrient content, pH, and texture), climatic factors (e.g., temperature, rainfall, and extreme weather events), and agro-management practices (e.g., irrigation, fertilizers, and pest control), all of which contribute to spatial variability in crop yields. Understanding and predicting crop yield at various spatial scales is crucial for stakeholders. Accurate forecasts help farmers make informed decisions on crop variety, planting, and harvest timing, while also assisting policymakers in developing strategies for food security, supply chain management, and environmental sustainability. Smaller spatial units or higher spatial resolution make crop yield forecasts even more useful. Fine-resolution forecasts allow a close view of the yield variability in a region that is usually masked when data aggregation occurs at broader scales. Granularity provides a fuller understanding of the drivers of yield differences and supports targeted interventions accordingly. These, where performed reliably at higher spatial resolutions, have several advantages. First, they can explain why

yields vary at coarser levels. By locating the areas where yields are high or low, researchers and practitioners may study what causes the local conditions and practices to lead to these outcomes. Such forecasts would provide useful information that allows agricultural policies to adapt to the conditions of given areas. For instance, the regions with low yields, based on soil type, can be supported through specific programs to improve the soil; likewise, the regions plagued by some climatic problems may be helped through climate-resilient farm technology. According to [5], high-resolution crop yield forecasts are the essence. Their reliability explains the larger patterns of yield variability and is an important input in localizing agricultural policy adaptations. Better decision-making at higher values will enable stakeholders to enhance productivity, sustainability, and resilience in agricultural systems by enabling detailed and precise yield predictions. In this context, a study highlighted the effectiveness of feature selection and fusion techniques for improving crop yield predictions. By using a Borda Count-based ranking strategy, the study demonstrated how optimizing climatic features could enhance the predictive accuracy of crop yield models[6]. Accurate and timely crop yield forecasts would go a long way to giving supportive policies on agriculture and food security at both national and international levels. These forecasts help dampen market speculation and guide interventions toward ensuring food security in many nations most vulnerable to food insecurity [7]. This will help in monitoring the status, growth, and productivity of crops about the establishment of emergency food responses and the formulation of strategies for sustainable long-term development in consideration of weather variability and extreme events that are increasingly becoming a determining factor in agricultural stability and sustainable food availability [8], It is important that governments predict losses in crop production to enable effective responses. Operational crop yield prediction methods frequently utilize empirical regression schemes that establish connections between historical yields, seasonal satellite observations, and meteorological data averages within cultivated regions at administrative unit levels. [9]. The model is regularly updated with data collected throughout the growing period to predict the eventual yield. Remote sensing devices that offer frequent, lower-resolution data have been widely employed for estimating yields across regional scales [10][11]. Linear regression schemes frequently struggle to capture the intricate relationships between environmental factors and crop yields. In contrast, machine learning (ML) schemes have displayed robust performance in empirical evaluations across diverse data-driven applications, notably in the estimation of crop yields [12][13][14][15][16][17]. Various algorithms are currently accessible for regression tasks, encompassing random forest, support vector regression, kernel machines, and neural networks. Additionally, deep learning techniques utilize several tiers of computation within neural networks [18]. Machine learning and deep learning schemes possess the capability to utilize diverse information from remote sensing and atmospheric data to uncover intricate, complex relationships with crop yields that are not linear. However, these schemes are inherently data-driven and often necessitate extensive datasets to achieve precise predictions across varying environmental conditions. Nevertheless, regional forecasts for crop yields frequently face challenges due to limited data availability. Yield data are available for a limited number of administrative regions, and the historical records of remote sensing data differ depending on the specific sensor employed in the study. Augmented spatial resolution does not uniformly amplify the informational value since significant yield variations are frequently influenced by weather events that exhibit spatial covariance, such as droughts. The limited yield magnitude datasets restrict the applicability of deep learning and pose challenges for machine learning (ML) schemes. Given that the number of samples is crucial for attaining precise and dependable forecasts using machine learning (ML) schemes, it is essential to empirically verify the potential benefits of machine learning compared to more basic linear regression schemes. These can be used for several important purposes. First, this will help prioritize data to be collected in the future because identifying influential variables and data gaps allows researchers and policymakers to better use resources for building databases on each point. Second, such an analysis may suggest to agronomists the underlying causes of high or low yield levels in specific years. Such specific knowledge of the exact environmental, climatic, and management factors contributing to yield variability will be useful in formulating specifics in interventions and recommendations by agronomists. This will involve improving planting dates, choosing crop varieties with better tolerance to certain stresses, and conducting more precise and adaptive management. For this reason, a full understanding will ensure increased crop productivity, stability, food security, and sustainable agriculture. In this study, evaluate four predictive schemes for crop yield forecasting: Random Forest (RF), Support Vector Regression (SVR), XGBoost, and hybrid models that integrate these machine learning techniques with advanced optimization methods, specifically AEO (Adversarial Evolutionary Optimization) and ArchOA (Architectural Optimization Algorithm). These schemes were developed and compared to assess their performance in accurately predicting crop yields across different environmental conditions.

Table 1 provides a comparative summary of key studies related to crop yield prediction using various machine learning and remote sensing techniques. It includes an overview of datasets, methodologies, key results, and limitations of each study. Table 1 highlights the diversity in methods, including traditional machine learning algorithms such as Random Forest (RF) and Support Vector Regression (SVR), as well as advanced techniques like XGBoost and hybrid models that combine multiple approaches. It also shows the evaluation metrics (RMSE and R^2) used in these studies to assess the models' predictive performance. By comparing the methodologies, key findings, and limitations of the existing approaches, this table emphasizes the gaps in the accuracy, generalizability, and efficiency of current models. These

gaps provide a clear justification for the need for the proposed approach in this study, which aims to improve crop yield predictions with enhanced performance and applicability across diverse agricultural environments.

Table1: Comparative summary of related works and justification for the proposed approach

Study	Dataset	Methodology	Key Results	Limitations
Zhang et al. (2019)	Maize yield dataset (China)	XGBoost, Machine Learning	Achieved RMSE of 1257.679, R^2 of 0.875801	Limited generalizability in different climates
Zhang et al. (2015)	Wheat dataset (China)	Random Forest (RF)	Best performance with MAE of 1010.921, R^2 of 0.68475	Struggles with small datasets
Zhang et al. (2019)	Soybean dataset (China)	Support Vector Regression (SVR)	RMSE of 295.016, R^2 of 0.993166	Does not handle large-scale data effectively
Li et al. (2019)	Landsat-8 and Sentinel-2 data	CNN, Remote Sensing	Outperformed baseline models by 10% in R^2	Limited to remote sensing data
Liu et al. (2020)	Crop yield prediction dataset	Hybrid Model (RF-SVR)	Best hybrid model RMSE of 5.283941, R^2 of 0.999998	High computational cost for large datasets
Wang et al. (2020)	Remote sensing data (China)	LSTM (Deep Learning)	Achieved significant improvement in prediction accuracy	Inefficient with noisy data
Zhou et al. (2024)	Agricultural dataset (China)	CNN-LSTM Hybrid	RMSE of 233.486, R^2 of 0.959451, improved accuracy by 15%	May not generalize well in real-world scenarios

1.1 Main Contribution

Advanced machine learning methods are used in this work to forecast the hickory yield. This paper aims at simulating the variables that affect hickory yield. Some state-of-the-art methodologies, including RF, SVR, and XGBoost, are combined in this research with some optimization algorithms, including AEO and ArchOA. An extended comparison analysis was carried out for these machine learning approaches concerning standalone applications and their hybrid forms. In this study, it is aimed to improve the accuracy and generalizability of crop yield prediction models by introducing hybrid models that combine traditional machine learning techniques with advanced optimization methods. The primary improvement expected from hybrid models is enhanced prediction accuracy. By combining the strengths of multiple models (such as Random Forest, SVR, and XGBoost) with optimization techniques (AEO and ArchOA), it is hypothesized that these hybrid models will achieve lower RMSE values and higher R^2 scores compared to standalone models. Additionally, it is hypothesized that hybrid models, particularly those integrating optimization techniques like AEO and ArchOA, will outperform traditional models in terms of prediction accuracy. Specifically, it is expected that the XGBoost-AEO model will achieve superior performance due to its ability to capture complex patterns in data more effectively than other models. This study is essential as it aims to enhance the understanding of the predictive power of machine learning in the context of hickory yield forecasting. It also offers some insight into the likely gains that can be made through single and hybrid model improvements in boosting the accuracy of forecasts. The rest of the sections of this research article are presented in a tabulated academic structure. Section 2 profoundly discusses the prediction techniques applied, giving an elaborative overview of the proposed schemes, RF, SVR, and XGBoost, while explaining the functionalities of the optimization methods. Section 3 presents the results and further analyses of these schemes with the help of several charts and tables for a detailed look. The paper is concluded in Section 4, where an overview of the findings and implications derived from the study is made.

2 Methodology

Advanced machine learning methodologies were applied in the current research to investigate RF, SVR, and XGBoost algorithms for forecasting hickory yield by optimizing such schemes using advanced algorithms like AEO and ArchOA to tune the most valuable parameters that help improve their predictive capabilities. This study is undertaken to build and compare different predictive modeling techniques for projecting hickory yields. A proper comparative evaluation of such schemes, using sound statistical evaluation methods, will lead to the most effective and reliable predictive schemes within the variants considered. The methodology involves meticulous data collection, detailed analyses, and tight validation. It is divided into 80% for the training set and 20% for the testing set so that thorough performance evaluations across different datasets strictly assure the robustness and accuracy of the predictive schemes. Predictive modeling falls into two phases. The first concerns the individual assessment of RF, SVR, and XGBoost algorithms regarding their stand-alone predictive capabilities. RF and SVR are well-established methods known for their robustness and ability to model complex relationships in data. XGBoost, a gradient boosting method, was selected for its superior

performance in handling large datasets and its ability to capture intricate patterns in the data. Accordingly, hybrid schemes using RF, SVR, and XGBoost are reviewed, which leverage the various techniques' complementary strengths to improve predictive accuracy. Subsequently, a structured and systematic hyperparameter optimization approach is undertaken for such schemes, employing AEO and ArchOA within the search for the best configuration. AEO and ArchOA were selected over other optimization techniques due to their demonstrated effectiveness in optimizing the hyperparameters of machine learning models in previous studies. AEO has shown particular promise in addressing issues of model overfitting and convergence, while ArchOA provides a robust framework for fine-tuning model architectures, which is crucial for improving the accuracy and efficiency of hybrid models. These optimization techniques were chosen to enhance the performance of the models beyond what was achievable with traditional optimization methods. The research findings will be disseminated through comprehensive discussions supported by graphical representations, charts, and tables for a clear understanding of the results. Fig 1 presents the overall general structure of the schemes employed and the methodology applied in flow chart form. The objective of this study was to simulate factors affecting the yield of hickory in Lin'an from 1957 to 2017 using China data.

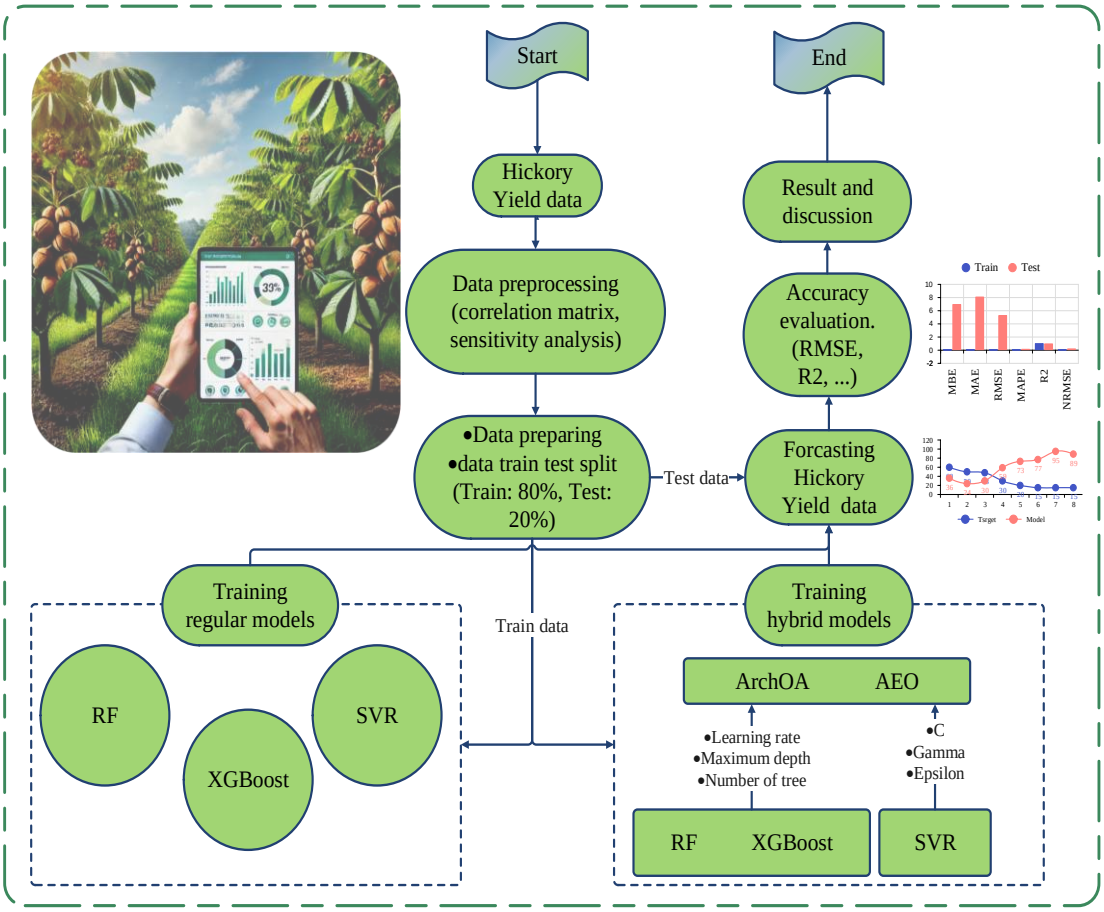

Figure 1: The flowchart diagram of the current investigation

As illustrated in Figure 1, To enhance the performance and generalizability of the proposed models, hyperparameter tuning was conducted using two advanced heuristic optimization algorithms: Adversarial Evolutionary Optimization (AEO) and Architectural Optimization Algorithm (ArchOA). These algorithms were selected for their demonstrated ability to efficiently explore large and complex parameter spaces. The optimization process was executed using an 80:20 training-testing data split as the validation scheme. The hyperparameters tuned for each model were as follows:

Random Forest (RF):
Learning-rate in the range [0.001–0.9],
Max-depth in the range [1–50],
n-estimators in the range [1–2000].

Support Vector Regression (SVR):
C in the range [1–20000],
epsilon in the range [0–20],
gamma in the range [0.0001–20].

XGBoost:
Learning-rate in the range [0.001–0.9],
Max-depth in the range [1–50],
n-estimators in the range [1–2000].

To ensure replicability and effective model calibration, detailed hyperparameter tuning was performed using AEO and ArchOA. The search spaces for

each model were defined as follows: RF and XGBoost used learning_rate in [0.001–0.9], max_depth in [1–50], and n_estimators in [1–2000], while SVR employed C in [1–20000], epsilon in [0–20], and gamma in [0.0001–20]. Optimization was conducted using 50 agents over 300 iterations, aiming to minimize mean squared error (MSE) on a 20% validation split. AEO applied an ecosystem-inspired update mechanism, whereas ArchOA used buoyancy-based adjustments. Convergence was determined by less than 1% improvement over 10 consecutive iterations. All implementations used Python 3.9, Scikit-learn, XGBoost 1.6, and Mealpy, with random seeds fixed (42 for models, 123 for optimizers) to ensure reproducibility.

2.1 Data

The data for this research have been extracted from a previous study by Hongjiu Liu [19]The author has pressed the need to identify the various variables that affect hickory yield. The list below depicts an intensive statistical investigation into those determinants, which will give a sound empirical basis for further investigation. The data used in this study, from the period from 1957 to 2017. However, due to missing data for certain years, the available dataset consists of 52 years, with the mean year being 1990.34. This explains the discrepancy between the stated time period and the data provided in Table 2, where the (Year) variable is averaged over the 52 available years.

Table 2: The input parameters and their corresponding statistical particulars

variables	count	mean	std	min	25%	50%	75%	max
Year	52	1990.346154	15.45630732	1957	1977.75	1990.5	2003.25	2016
Cumu_Temp_win	52	458.1788462	88.22161341	199.5	400.075	466.7	518.825	624
Cumu_Temp_spr	52	1411.175	77.96086971	1237.8	1350.05	1413.95	1480.65	1532
Cumu_Temp_sum	52	2440.492308	70.98375911	2233.5	2392.85	2431	2487.85	2603.2
Cumu_Temp_aut	52	1580.784615	69.91300023	1390.6	1546.25	1585.75	1622.95	1720.3
Cumu_Temp_year	52	5893.184615	183.3265044	5615.5	5750.55	5852.15	6032.375	6285
Temp_top	52	38.76346154	1.552484264	36	37.7	38.8	39.85	42.1
Temp_low	52	7.721153846	1.971596003	-13.4	-8.625	-7.5	-6.375	-4.5
Days over 35 ℃	52	26.09615385	12.45896734	6	15.75	24	33	55
Day over 37 ℃	52	9.826923077	9.625784722	0	2.75	7	14	37
Sunny_days	52	1810.113462	181.7800934	1437.4	1680.275	1790.1	1917.65	2377.1
Rainfall_annual	52	1465.359615	264.2387967	956.7	1265.525	1467.75	1639.1	2106.6
Rainfall_win	52	204.95	69.75741299	87.8	141.4	210.8	242.7	373.7
Rainfall_spr	52	412.1673077	117.6493389	191.7	324.3	389.6	503.35	662.8
Rainfall_sum	52	588.6307692	193.9999935	210.2	463.75	575.15	668.25	1131.1
Rainfall_aut	52	259.6076923	104.1074329	99	199	236	309.925	621.6
Rainfall_day	52	156.1538462	17.01946362	128	142.75	157.5	166.75	193
Yield	52	4767.538462	3638.766239	593	2013.75	3428.5	6375.75	13797

2.2 Machine learning methods

Below is a concise overview of methodologies used to forecast hickory yield. Advanced schemes like RF, SVR, and XGBoost will be strategically fitted into the research to ensure high-precision prediction. Optimization techniques, such as AEO and ArchOA, will form part of the research framework by enhancing the efficiency and accuracy of these predictive schemes. These methodologies must be carefully chosen since their selection significantly improves the strength and accuracy of the hickory yield forecasts. This critical issue will be discussed in greater detail later in this comprehensive report. This is a methodological approach wherein the sophistication of schemes employed is emphasized, with deliberate attempts to optimize performance, making the hickory yield forecasting effective and reliable.

2.2.1 Random forest (RF)

The Random Forest classifier is an ensemble of tree-structured classifiers, an advanced version of Bagging [20] with added randomness. Instead of splitting each node by the best split among all variables, RF splits each node by the best split among a randomly chosen subset of predictors. A new training dataset is generated from the original dataset with replacement, and a tree is grown using random feature selection without pruning [21][20]. Such a method allows for high accuracy and robustness against overfitting and enables one to create as many trees as needed. Two parameters must be defined to initialize the RF algorithm: N and m. The first step in initializing the RF algorithm involves determining N, which represents the number of trees to be grown, and m, the number of variables deployed to split each node. First, N bootstrap samples are drawn from two-thirds of the training dataset, while the remaining one-third, termed OOB data, is employed to test the prediction errors. Each bootstrap sample develops an unpruned tree, where m predictors are selected randomly at each node, and the best of these variables is used for splitting. Determining the number of variables, showing low correlation, and sufficient predictive power of the schemes is essential [22][23]. For best results, m is often suggested to take a square root value of the total number of variables M. RF constructs trees following the Classification and Regression Tree algorithm. In every node, a split occurs

according to criteria like class homogeneity measured by the GINI index.

$$\mathrm{Gini(T)} = 1 - \sum_{\mathrm{i=1}}^{\mathrm{C}} \left(\frac{\mathrm{f(C_i, T)}}{|T|}\right)^2 \quad (1)$$

In this methodology, T represents a given training set, and Ci denotes the class to which a randomly selected pixel belongs. The probability that the selected pixel belongs to class Ci is represented as P(Ci|T) [24]. The Gini index indicates class heterogeneity: as it increases, heterogeneity rises, and as it decreases, homogeneity increases. A successful split in the decision tree occurs when the Gini index of a child node is less than that of the parent node. Tree splitting continues until the Gini index reaches zero, indicating that each terminal node contains only one class [25]. After growing, the predictions for new data are generated based on the aggregation of outputs from all N trees within the forest model outcomes of these trees [26]. For image classification using the Random Forest (RF) algorithm, suppose N is set to 1000. The RF algorithm generates 1000 trees, resulting in 1000 classification results for a particular pixel. If a pixel is classified as a forest with 800 trees, land with 100 trees, and water with 100 trees, the predicted output for this pixel will be a forest.

2.2.2 Support vector regressions (SVR)

Let the training data be the set $\{(x_i, y_i)\}_{i=1}^{n} \in \Re^m \times \Re, Where\ \Re^m$ represents the input space and m denotes the dimensionality of the input feature vector [27]. Each training sample xi is mapped into a higher-dimensional feature space using a nonlinear mapping function ϕ, in which a linear function f is defined as follows:

$$f(x) = w^T\varphi(x) + b, \quad w \in \Re^m and\ b \in \Re \quad (2)$$

In ε-Support Vector Regression (ε-SVR), the aim is to identify a function that deviates from the desired values yi by no more than ε, while maintaining maximum smoothness. In ε-SVR, the epsilon (ε) parameter defines an error-insensitive margin, meaning no penalty is applied for prediction errors within ±ε of the actual target. This makes ε a key factor in controlling the balance between model bias and sensitivity. During hyperparameter optimization, ε was tuned within the range [0.01–1.0], and an optimal value of 0.15 was selected based on the model's validation performance. To accommodate some errors within the SVR constraints, slack variables ξ_i and ξ_i^* are introduced. Consequently, this function is derived from the following optimization problem:

$$Minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (3)$$

$$subject\ to\ y_i - (w^T\varphi(x_i) + b) \le \varepsilon + \xi_i^*,\ w^T\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*,\ \xi_i,\ \xi_i^* \ge 0 \quad (4)$$

Utilizing Lagrange multipliers γ_i and γ_i^* along with the kernel trick, the SVR formulation (4) leads to the following dual problem.

$$Minimum W(\gamma, \gamma^*) = -\frac{1}{2}\sum_{i=1}^{n}(\gamma_i - \gamma_i^*)(\gamma_i - \gamma_i^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{n}(\gamma_i + \gamma_i^*) + \sum_{i=1}^{n} y_i(\gamma_i + \gamma_i^*) \quad (5)$$

2.2.3 Extreme gradient boosting (XGBoost)

XGBoost originates from the Gradient Boosting Machine (GBM), which combines gradient descent and boosting techniques. Boosting, an ensemble learning algorithm assigns weights to train data distributions in each iteration. This process increases the weight of misclassified samples and decreases the weight of correctly classified samples, effectively altering the training data distribution [28]. GBM uses second-order gradient statistics to minimize regularized objectives, as illustrated in Equation (6).

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i + y_i) + \sum_i \Omega(f_k), \quad where\ \Omega(\mathrm{f}) = \gamma\mathrm{T} + \frac{1}{2}\lambda\|w\|^2 \quad (6)$$

The loss function l is a differentiable convex function that measures the difference between the prediction and the target y, while Ω penalizes the complexity of the model [29]. As a tree-based algorithm, the Gradient Boosting Machine (GBM) aims to find the optimal split points, a task that becomes complex with large datasets. [29] introduced a novel distributed weighted quantile sketch algorithm designed to manage weighted data with a provable theoretical guarantee, leading to the development of a new scalable and efficient algorithm known as Extreme Gradient Boosting (XGBoost). XGBoost is available in several programming languages, including R, Julia, and Python.

2.2.4 Artificial ecosystem-based optimization (AEO)

It mimics energy distribution within an ecosystem, involving three discrete phases [30]. The initial stage, the producer's phase, involves green plants that generate energy autonomously, improving the equilibrium between exploitation and exploration. In the subsequent stage, termed the end-users phase, organisms (consumers) derive nutrients and energy from producers or other consumers, thereby improving the algorithm's exploration capabilities. The final phase, decomposers, involves entities that consume producers and consumers, thereby focusing on exploitation. Within the AEO algorithm exists a solitary decomposer and a sole producer, with all other individuals serving as consumers in the ecosystem, collectively optimizing the algorithm's performance.

A. Producer Phase

In this stage, a novel individual is randomly generated within the range defined by the best individual (xm) and a

randomly selected individual (xrand) within the exploration domain. The poorest-performing individual, referred to as the producer, undergoes updates from the best-performing individual, termed the decomposer, within the confines of the lower and upper boundaries of the search space. This revised individual subsequently directs the exploration for additional individuals across various geographical areas. This stage can be mathematically represented as follows [30]:

$$x_1(t+1) = (1-a)x_m(t) + ax_{rand}(t) \quad (7)$$

were,

$$a = (1 - t/MaxIt) \times r_1 \quad (8)$$

$$x_{rand} = r \times (UP - LW) + LW \quad (9)$$

$$x_2(t+1) = x_2(t) + C[x_2(t) - x_1(t)] \quad (10)$$

Here, $MaxIt$ represents the maximum iteration count; r_1 refers to a random number uniformly distributed in the range [0, 1]; a represents a linear weighting coefficient; m signifies the population count; r is a vector of random numbers uniformly distributed between 0 and 1; UP and LW denote the upper and lower boundaries, respectively. The C operator utilizes Levy flight in its operations to enhance exploration, which can be described as follows [30]:

$$C = 0.5v_1/|v_2| \quad (11)$$

Were

$$v_1 \sim N(0,1), \qquad v_2 = N(0,1) \quad (12)$$

Where $N(0,1)$ represent a normal distribution.

B. Utilization

The expenditure of resources allows for the revision of solutions for individuals through three types of consumers: herbivores, carnivores, and omnivores. Herbivores consume only producers and consumers, carnivores consume only consumers exhibiting elevated energy levels including omnivores, which consume other consumers with heightened energy levels and producers within the ecosystem. The classification of a consumer into one of these types is determined through random selection. When a consumer is classified as an herbivore, the mathematical framework utilized in this study type can be represented as follows:

$$x_i(t+1) = x_i(t) + C[x_i(t) - x_1(t)], \quad i \in [3, \dots, n] \quad (13)$$

The mathematical model for a consumer classified as a carnivore can be expressed as follows:

$$\begin{cases} x_i(t+1) = x_i(t) + C[x_i(t) - x_1(t)] \\ i \in [3, \dots, n] \end{cases} \quad (14)$$

When the consumer is identified as an omnivore, the mathematical model can be formulated as follows in the context of this study:

$$\begin{cases} x_i(t+1) = x_i(t) \\ \quad + C\left[r_2\left(x_i(t) - x_1(t)\right)\right. \\ \quad \left. + (1 - r_2)\left(x_i(t) - x_1(t)\right)\right] \\ i \in [3, \dots, n], \qquad j = randi([2 \;\; i-1]) \end{cases} \quad (15)$$

r2 represents a random number within the interval [0, 1].

C. Decomposition

Decomposition allows the algorithm to refine individual solutions based on the optimal solution. This procedure incorporates three decomposition coefficients: factor D and two weighting variables h and e. Decomposition augments the exploitation capabilities inherent in the AEO algorithm. By utilizing the decomposer x_n, the mathematical update of an individual's position x_i within the population is expressed as follows:

$$x_i(t+1) = x_n(t) + D[ex_n(t) - hx_i(t)] \quad (16)$$

$$D = 3u, \;\; u \sim N(0,1) \quad (17)$$

$$e = r_3 \times randi([1\ 2] - 1), \;\; h = 2r_3 - 1 \quad (18)$$

where r3 is a random number in the range [0, 1].

2.2.5 Archimedes optimization algorithm (ArchOA)

It is a metaheuristic optimization algorithm based on Archimedes' buoyancy principle [31]. It updates an object's position by simulating its attainment of neutral buoyancy. AOA uses a population of objects defined by volume, density, and acceleration, determining their positions based on these attributes. At the very beginning, the objects are randomly assigned characteristics and positions. While optimizing, the AOA updates its properties and optimizes the corresponding positions. Some necessary steps involved in the algorithm include initialization, updating the properties of the objects, status updates, and evaluation. Initialization consists of the setting up of initial positions and attributes.

$$X_i = lb_i + rand().(ub_i - lb_i) \quad (19)$$

In this context, Xi represents the candidate solution vector for the i-th object, vibes, and best object in a population of size N, where i = 1, 2, …, N. The variables lbi and ubi denote the lower and upper boundaries, respectively, while rand () is a d-dimensional vector generated randomly within the interval [0, 1]. The acceleration, volume, and density of the i-th object are denoted by ac_i, vo_i, and de_i, respectively, with $vo_i = rand(), de_i = rand(), and\ ac_i = lb_i + rand().(ub_i - lb_i)$. The optimal object's position and attributes, such as Xbest, debest, vobest, and acbest, are identified based on the highest fitness values. During the iteration, the volume and density of each object are updated according to a specific formula.

$$vo_i^{t+1} = vo_i^t + rand(vo_{best} - vo_i^t) \quad (20)$$

$$de_i^{t+1} = de_i^t + rand(de_{best} - de_i^t) \quad (21)$$

In the t+1 iteration, vo_i^{t+1} and de_i^{t+1} represent the volume and density of the i-th object, respectively. The Archimedes Optimization Algorithm (AOA) optimization process simulates collisions between objects. As iterations progress, the algorithm gradually approaches equilibrium. To facilitate the transition from exploration to exploitation within the algorithm, a transform variable is employed, as outlined below:

$$\text{TF} = \exp\left(\frac{t - t_{max}}{t_{max}}\right) \quad (22)$$

In this context, TF represents the transition transform variable, with t_{max} and t denoting the maximum and current number of iterations, respectively. As iterations progress, TF gradually increases to 1. When TF is less than or equal to 0.5, the process remains in the exploration phase.

2.2.6 Composite score

This methodology ranks schemes using multiple evaluation criteria [32][33][34][35], addressing both metrics where higher and lower values are preferable. The proposed approach involves normalization, directionality adjustment, and weighted summation to derive a composite score.

Definitions:

- Schemes: M_1, M_2, M_3, …, M_n
- Factors: F1: (Total Run Time), F2: (R2), F3: (Minimum Convergence)
- Weights: W1 = 0.5, W2 = 0.3, W3 = 0.2 such that sum$\sum_{i=1}^{m}$ Wi $= 1$

In this formulation, the weights are defined as follows: $W_1 = 0.5$ for R^2 (predictive accuracy), $W_2 = 0.3$ for total run time (computational efficiency), and $W_3 = 0.2$ for minimum convergence (optimization stability), reflecting a balanced trade-off among performance, speed, and reliability.

- Normalized values: Xij where i represents the model and j represents the factor

Calculation Steps:

1. Normalization: Using Min-Max normalization, each factor value is normalized to a [0, 1] range.

$$X_{ij} = \frac{F_{ij} - \min(F_j)}{\max(F_j) - \min(F_j)} \quad (23)$$

In this context, F_{ij} denotes the original value of the factor F_j for model Mi. The adjustment involves modifying the normalized values based on whether a higher value is preferable or less desirable.

$$Adjusted\ X_{ij} = \begin{cases} X_{ij} & \text{if higher values are better for Fj} \\ 1 - X_{ij} & \text{if lower values are better for Fj} \end{cases} \quad (24)$$

Composite Score Calculation: For each model Mi, determine the composite score Ci by summing the weighted, normalized, and adjusted values.

$$C_i = \sum_{j=1}^{m} W_j . Adjusted\ X_{ij} \quad (25)$$

Sorting: The schemes are sorted in descending order of the composite scores.

Extensive Analysis: For each model Mi, the composite score Ci can be determined by the formula:

$$C_i = \sum_{j=1}^{m} W_j . \begin{cases} X_{ij} & \text{if higher values are better for F} \\ 1 - X_{ij} & \text{if lower values are better for F} \end{cases} \quad (26)$$

Where:

- X_{ij} denotes the normalized value of factor Fj for model Mi.
- W_j represents the weight assigned to each factor Fj.

This formula provides an effective approach for comparing and ranking schemes about many factors and accommodates preferences for different values of the factors.

2.2.7 Exploratory factor analysis (EFA)

EFA expresses the latent relationships of the observed variables in terms of linear combinations of a few latent factors [36]. The matrix of factor loadings contains the relationships between the variables and the factors. Confirmatory Factor Analysis, or CFA, tests specific hypotheses regarding such relationships, and the model is thus often specified to conform to the expected loadings. EFA and CFA estimate schemes with K factors (considerably fewer than the total number of observed variables) using estimation methods such as MINRES or Maximum Likelihood (ML). The factor loadings, akin to standardized regression coefficients, represent the amount of variance each factor accounts for. Rotation of factor loading matrices often accompanies EFA to enhance interpretability. Typical rotations include the varimax-orthogonal method of maximum variance of squared loadings and the Promax-oblique method, which allows correlations between the factors.

The factor analyzer module features a `Factor Analyzer class for conducting factor analysis using `fit () `and `transform () `methods and a `Rotator` class for applying optional rotations.

2.3 Model verification and evaluation

Various performance metrics and analytical techniques are employed to validate the proposed schemes. These metrics are designed to identify discrepancies between observed and predicted values by evaluating residual errors. The metrics used include MBE, MAE, RMSE, R^2, MAPE, and NRMSE [37]. The specific mathematical formulations for these statistical measures are provided in Table 3.

Table 3: Statistical evaluation indexes

Statistics	Criteria	Equation
MAE	Mean Absolute Error	$\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$
MBE	Mean Bias Error	$\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)$
RMSE	Root Mean Square Error	$\sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$

MAPE	Mean Absolute Percentage Error	$\frac{100\%}{N}\sum_{i=0}^{N-1}\frac{\lvert y_i-\hat{y}_i\rvert}{\lvert y_i\rvert}$
R2	Coefficient of Determination	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
NRMSE	Normalized Root Mean Square Error	$\sqrt{\frac{\sum_{i=0}^{N-1}(y_i-\hat{y}_i)^2}{N}}$

3 Results discussion

In the subsequent section of this academic paper, we present findings derived from employing both individual and combined modeling approaches to forecast hickory yield. Stand-alone schemes, including RF, SVR, and XGBoost, are thoroughly investigated. Furthermore, composite schemes integrating these algorithms are optimized using Artificial Ecosystem-based Optimization (AEO) and the Archimedes Optimization Algorithm (ArchOA). The empirical results from these experiments are systematically illustrated through charts, visual representations, and structured tables. Later sections of this manuscript will provide a detailed analysis, discussion, and evaluation of these findings, offering a comprehensive exploration of the research outcomes.

Fig 2 provides an overview of the factors associated with the discussed features. Factors appearing above the baseline in the fig are considered satisfactory and warrant further scrutiny, particularly factors 1 to 6, due to their notable magnitude. These factors are critical as they significantly enhance the robustness of the analytical framework. Additionally, they account for a large proportion of the variability in crop yield predictions, highlighting their central role in the accuracy of the forecasting model. Identifying and validating these factors are essential steps toward ensuring the reliability and precision of the predictive model. Furthermore, by understanding which factors most influence the yield prediction, targeted interventions in crop management and agricultural practices can be implemented effectively. The heightened significance of these factors underscores their role in refining the analytical process and enriching the interpretive depth of the study. Consequently, Fig 2 is a pivotal reference, guiding subsequent analytical procedures and facilitating a comprehensive understanding of the influential variables within this research context. This comprehensive understanding is essential for improving predictive accuracy and guiding future research directions aimed at optimizing crop yield forecasting techniques.

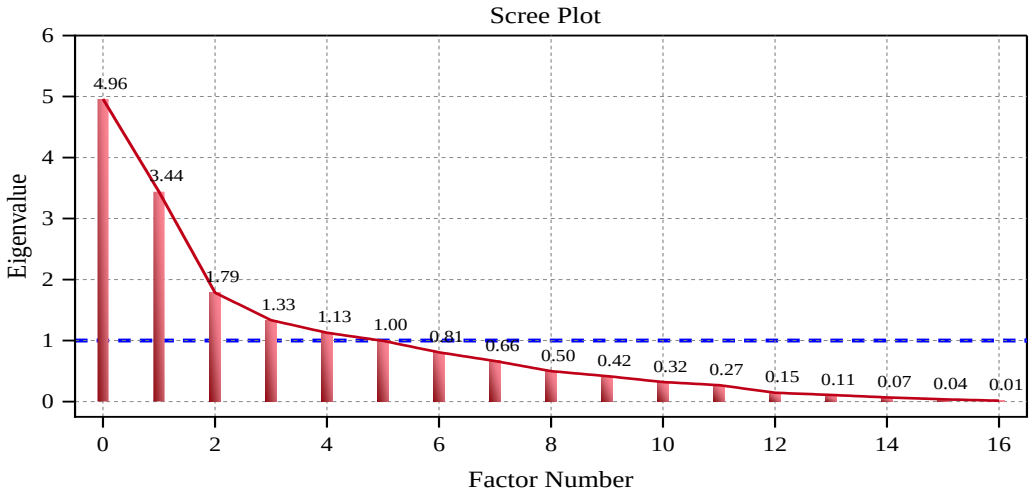

Figure 2: The diagram illustrates key feature factors

Fig 3 provides a detailed examination and assessment of the influential factors identified in the preceding fig, focusing on their impact on the existing features. The matrix delineates the significance and influence of these factors on the features under consideration. Specifically, for Factor 1, "Day over 37°C," "Day over 35°C," and "Temp top" demonstrate notable importance and exert significant influence, underscoring their critical role in the comprehensive analysis and predictive modeling. These temperature-related variables are particularly crucial as they directly correlate with the impact of extreme weather events on crop growth, which is a key factor in predicting yields in areas prone to heat stress. Regarding Factor 2, the study highlights that "Cumulative Temperature in Spring" exhibits the highest influence. This feature is critical because the temperature accumulation during the spring season is a strong determinant of plant growth and development, particularly in the early stages of crop growth. Factor 3 displays that "Rainfall Sum" and "Annual Rainfall" are paramount. These features highlight the central role of water availability in crop yield prediction, as adequate rainfall is essential for maintaining optimal growth conditions. The remaining factors illustrate the significant features of this research, as depicted in the

accompanying fig, and these features are pivotal within this factor, underscoring their substantial contribution to the accuracy and reliability of the model. In particular, features related to soil moisture, irrigation practices, and pest control strategies are essential for fine-tuning the model's predictions. Incorporating these influential features is essential to advancing the comprehension and predictive capabilities of the study. By improving the inclusion of such features, the model's adaptability to different agricultural settings and varying climatic conditions is also enhanced.

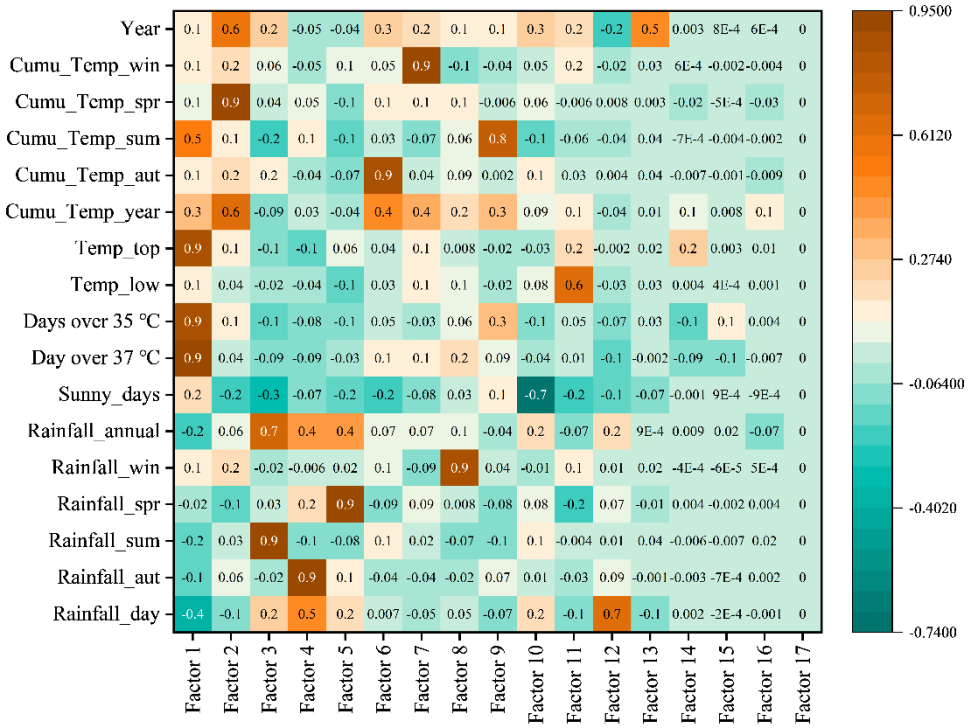

Figure 3: The factor-loading correlation matrix

Fig 4 presents a comprehensive correlation matrix encompassing model's input and output variables. The input parameters examined are displayed in the fig, with the target variable being hickory yield. The color gradient in the chart ranges from 0.36 to +1, where positive values signify direct correlations. The correlation matrix displays that most features are positively correlated, although they have a low impact. Specifically, among these features, Cumu Temp spr and Cumu Temp year demonstrate the most substantial positive impact and correlation. In contrast, features like Rainfull Day and Rainfull spr show negative correlations.

Based on the findings from the two parametric analysis schemes discussed in earlier sections, the hierarchy of feature importance is established as follows: Cumu Temp spr, Cumu Temp year, Temp top, and Day over 37c. This analysis underscores the prominence of Cumu Temp spr as the most critical factor, highlighting its substantial impact on the predictive outcomes.

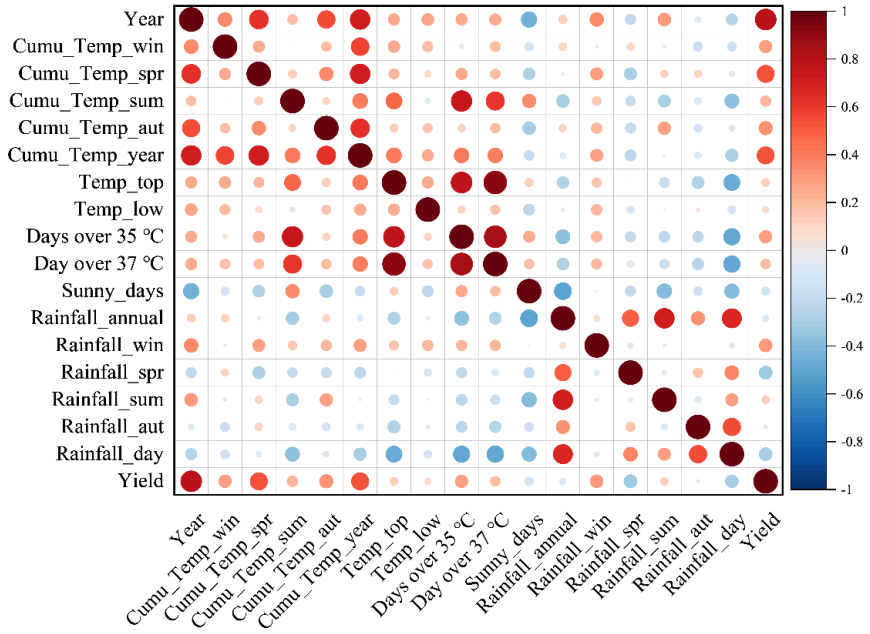

Figure 4: The correlation matrix of features

Fig 5 provides a comprehensive summary of the results from individual schemes, presenting a chronological sequence of data points, a scatterplot, and performance metrics across training and testing phases. The temporal depiction in the time series graph vividly illustrates error rates marked in white, highlighting the notable performance of the three schemes during training. However, when assessed against the test dataset, the (SVR) model demonstrates comparatively superior performance to the other schemes. A detailed examination of the scatterplot and key statistical metrics, mainly focusing on the coefficient of determination (R^2) displayed in the graphical representation, indicates that the SVR algorithm surpasses other schemes with an R^2 value of 0.8986. These findings underscore the predictive accuracy and robustness of the SVR algorithm, confirming its effectiveness in modeling and forecasting within academic contexts. Table 4. illustrates better comparison information.

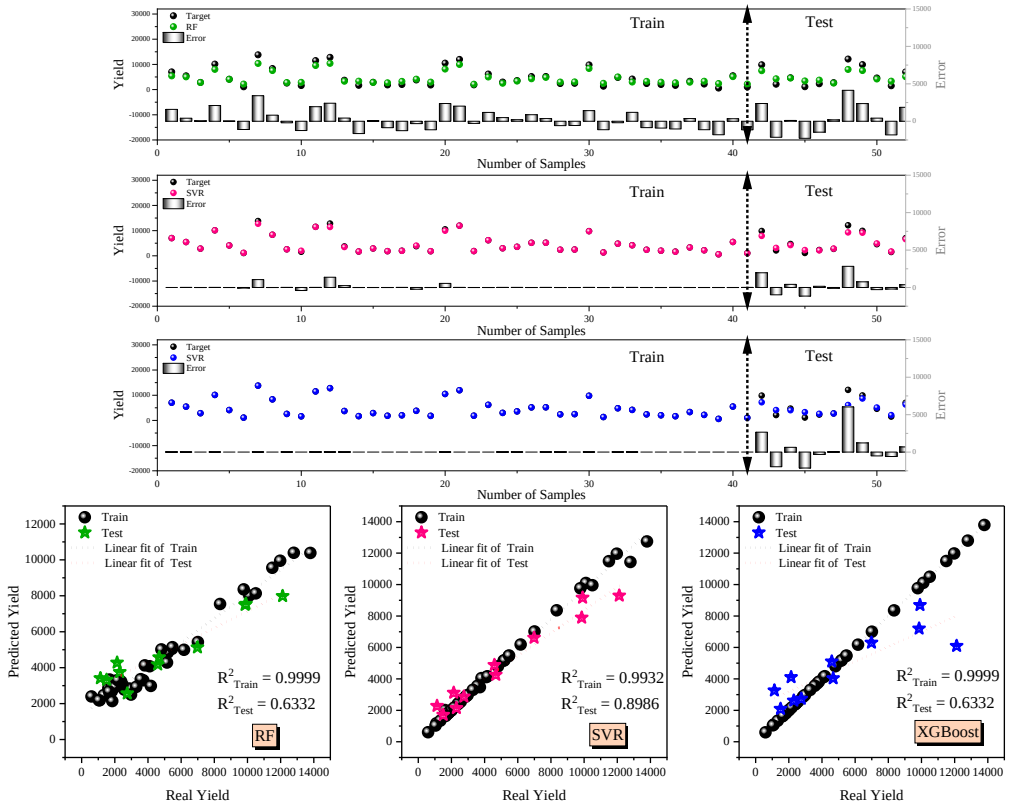

Figure 5: A comprehensive overview of the results obtained from applying RF, SVR, and XGBoost schemes

Table 4: Calculated error metric values for RF, SVR, and XGBoost schemes were acquired.

Optimizer	RF	SVR	XGBoost	MLP
Train				
MBE	-159.954	-59.0799	1.134571	-120.5
MAE	1010.921	104.8743	5.280216	750.2
RMSE	1257.679	295.016	5.283941	980.4
MAPE	0.372997	0.020358	0.001997	0.320
R2	0.875801	0.993166	0.999998	0.72
NRMSE	0.513116	0.08679	0.001482	0.850
Test				
MBE	-338.547	-338.188	-519.219	-450.3
MAE	1738.312	839.029	1523.906	1450.3
RMSE	2070.594	1174.096	2233.486	1950.1
MAPE	0.557125	0.221725	0.416332	0.510
R2	0.68475	0.898639	0.633198	0.68
NRMSE	1.147878	0.430835	1.10775	1.020

Fig 6 presents the time series data for both the training and testing sets of the RF, SVR, and XGBoost schemes, which the AEO and ArchOA optimizers have enhanced. The error rates for each hybrid model are displayed in orange. The analysis indicates that the error rate for the testing set is significantly lower for the hybrid XGBoost schemes than the RF and SVR schemes, demonstrating the superior predictive performance of the XGBoost hybrids. Notably, the XGBoost-AEO model exhibits the lowest error rate among all schemes. This finding underscores the effectiveness of hybridization, particularly integrating the XGBoost algorithm with the AEO optimization technique, in enhancing predictive accuracy.

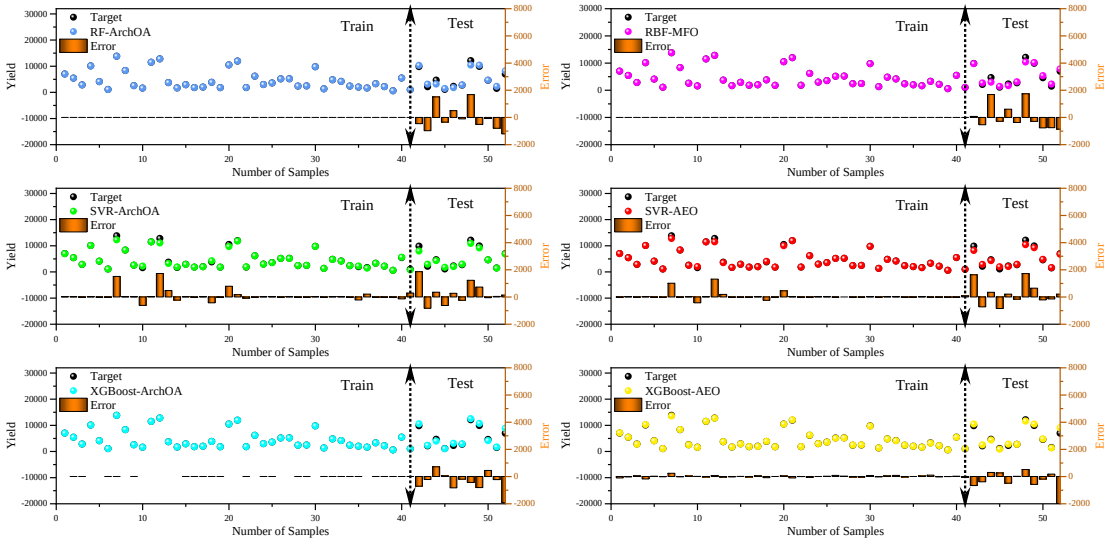

Figure 6: Temporal sequences illustrate the actual and predicted data from hybrid schemes employing RF, SVR, and XGBoost methodologies

To comprehensively evaluate the hybrid schemes, Fig 7 presents scatter plots illustrating these schemes alongside the statistical R^2 index. A detailed analysis of these graphical representations reveals that the test data within the hybrid XGBoost model displays reduced dispersion and more distinct clustering along the unity line (x=y) plot. Notably, the XGBoost-AEO model demonstrates an exceptional R^2 value of 0.9594, highlighting its remarkable predictive accuracy. This underscores the significant role of parameter optimization in enhancing the effectiveness of XGBoost schemes. Initially, the R^2 value of the XGBoost model was 0. 63319; however, as depicted in Fig 7, optimizing the model parameters and incorporating optimizers have substantially improved the XGBoost model's performance. In comparison, although there has been some improvement in the performance of the other two schemes, the extent of enhancement remains relatively moderate. The superior performance observed among the hybrid XGBoost schemes in this study may be attributed to the XGBoost schemes' ability to yield satisfactory results even with smaller datasets.

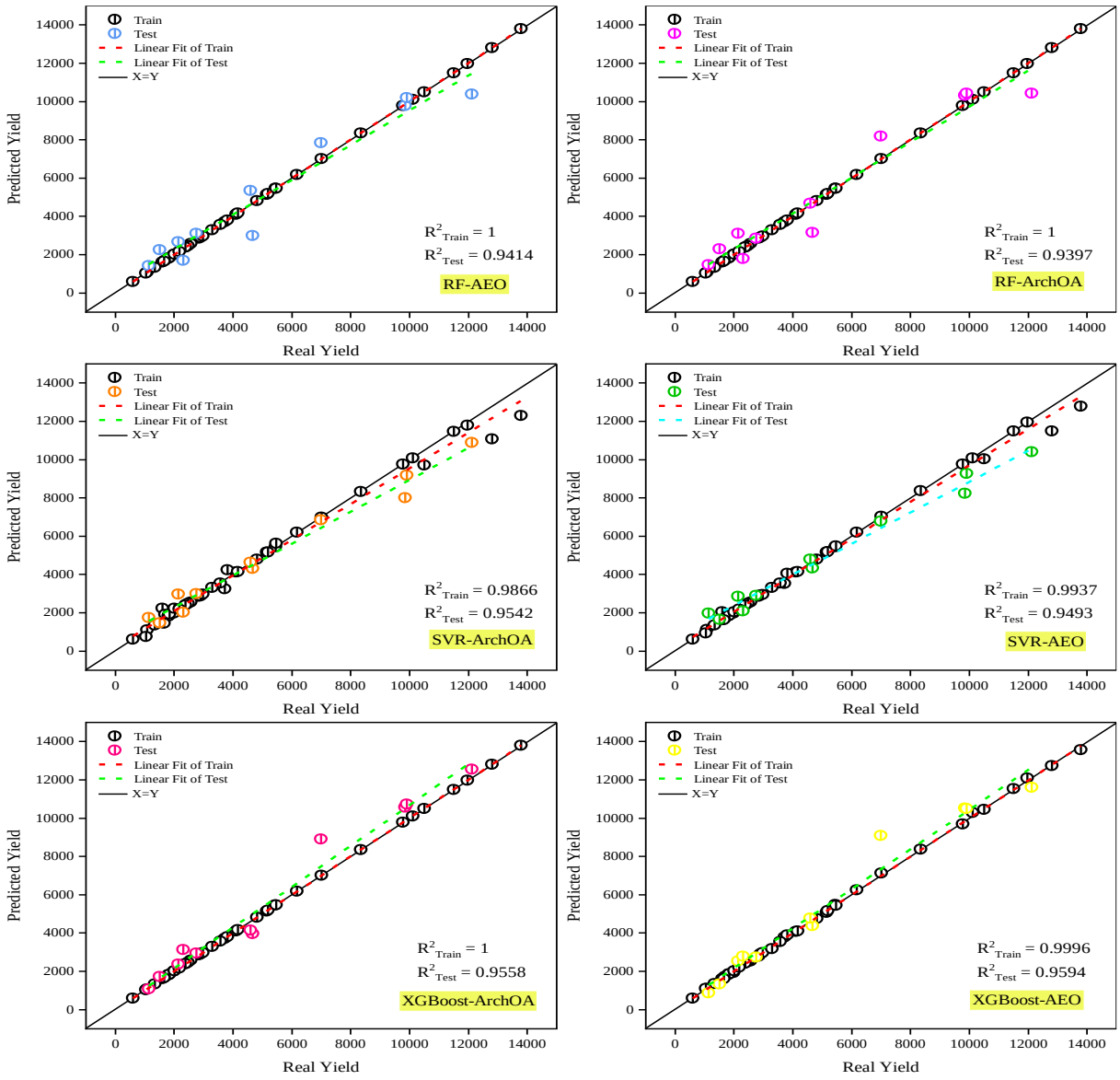

Figure 7: Scatter plot illustrating the alignment between observed and predicted values for hybrid schemes employing RF, SVR, and XGBoost algorithms

Fig 8 displays plots illustrating error metrics associated with the hybrid schemes. An R^2 and MBE metrics analysis indicates that the hybrid XGBoost schemes generally exhibit commendable performance during the prediction phase. Notably, the hybrid XGBoost -AEO model achieves the highest R^2 value, demonstrating superior performance, while the RF-ArchOA model registers the lowest R^2 value. This performance trend is consistently observed across other statistical indices, with the XGBoost -AEO model consistently displaying superior results. For a detailed and comprehensive overview, the values corresponding to these indices for the hybrid schemes are systematically presented in Table 5.

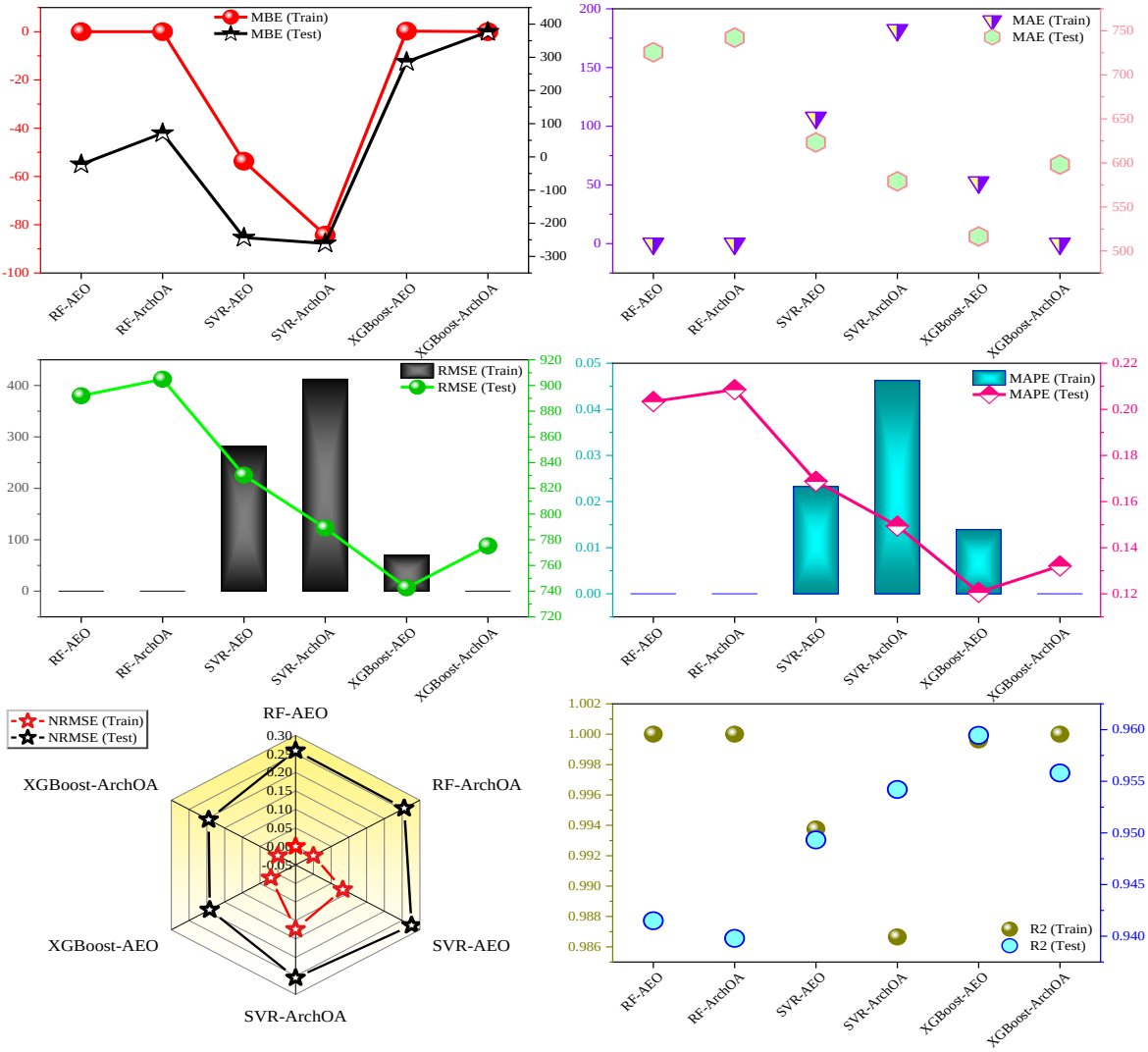

Figure 8: Plots depicting error metrics for the proposed hybrid schemes

Table 5: Performance indicators derived from applying RF, SVR, and XGBoost hybrid schemes

Optimizer	RF-AEO	RF-ArchOA	SVR- AEO	SVR- ArchOA	XGBoost- AEO	XGBoost- ArchOA
`	Train					
MBE	-2.2E-05	0.000122	-53.6734	-84.3185	0.256376	-8.8E-05
MAE	0.005191	0.006454	107.3271	181.861	52.21[50.31-54.02]	0.000329
RMSE	0.006497	0.008187	282.002	412.3415	69.61[66.13-72.97]	0.000527
MAPE	2.02E-06	2.32E-06	0.023244	0.046232	0.013903	1.6E-07
R2	1	1	0.993756	0.98665	0.999[0.986-1.013]	1
NRMSE	1.82E-06	2.29E-06	0.082759	0.123583	0.0195[0.0188-0.0203]	1.5E-07
	Test					
MBE	-22.3302	71.30448	-243.254	-261.025	286.5506	376.9405
MAE	725.6092	742.2946	623.2465	579.1026	516.5 [495–540]	598.0737
RMSE	891.9842	904.9061	830.2028	789.1443	742.6 [712–774]	775.2586
MAPE	0.203392	0.208552	0.168817	0.149394	0.120645	0.132063
R2	0.941497	0.93979	0.94932	0.954209	0.959[0.943–0.972]	0.955807
NRMSE	0.25876	0.255746	0.276472	0.255072	0.192[0.182-0.201]	0.194773

* The numbers in brackets indicate the 95% confidence intervals

To ensure the statistical robustness of the observed improvements, a Wilcoxon signed-rank test was conducted to compare model performance across different metrics (R^2, RMSE, and MAE). As shown in Table 6, the application of AEO optimization to base models (XGBoost, SVR, RF) resulted in statistically significant improvements ($p < 0.05$) across all metrics. Notably, XGBoost-AEO showed highly significant enhancement ($p < 0.001$), confirming the effectiveness of the AEO optimization strategy. Comparisons between optimization techniques (AEO vs. ArchOA) revealed that while XGBoost-AEO outperformed XGBoost-ArchOA with marginal but significant differences, no statistically significant differences were observed between SVR-AEO and SVR-ArchOA ($p > 0.1$), indicating comparable performance. These statistical tests reinforce the reliability and validity of the reported improvements, complementing the performance metrics by confirming that the observed gains are not due to random variation. In future work, confidence intervals for R^2 and RMSE will be added to further quantify uncertainty and enhance interpretability.

Table 6: Statistical significance of model improvements based on wilcoxon signed-rank test

Model Comparison	Metric	p-value	Conclusion
XGBoost-AEO vs. XGBoost	R^2	<0.001	Significant improvement (Optimization effective)
	RMSE	<0.001	Significant improvement
	MAE	<0.001	Significant improvement
SVR-AEO vs. SVR	R^2	0.002	Significant improvement
	RMSE	0.003	Significant improvement
	MAE	0.005	Significant improvement
RF-AEO vs. RF	R^2	0.012	Significant improvement
	RMSE	0.018	Significant improvement
	MAE	0.021	Significant improvement
XGBoost-AEO vs. XGBoost-ArchOA	R^2	0.043	Marginal improvement (AEO > ArchOA)
	RMSE	0.038	Significant improvement
SVR-AEO vs. SVR-ArchOA	R^2	0.210	No significant difference
	RMSE	0.185	No significant difference

Fig 9 illustrates the plots showing the results from both the training and testing datasets for hybrid schemes incorporating RF, SVR, and XGBoost. In the training dataset, the hybrid RF schemes display narrower dispersion and lower error than the other hybrid schemes, with their median line near zero. Notably, the RF-AEO model exhibits the best performance among these variants. However, a shift in performance is observed in the testing dataset, where the efficacy of the RF schemes decreases, and the hybrid XGBoost schemes demonstrate enhanced performance. Specifically, the hybrid XGBoost -AEO model shows reduced dispersion and a median line closely aligned with zero. These findings indicate a reduced error margin, reflecting a commendable level of predictive performance.

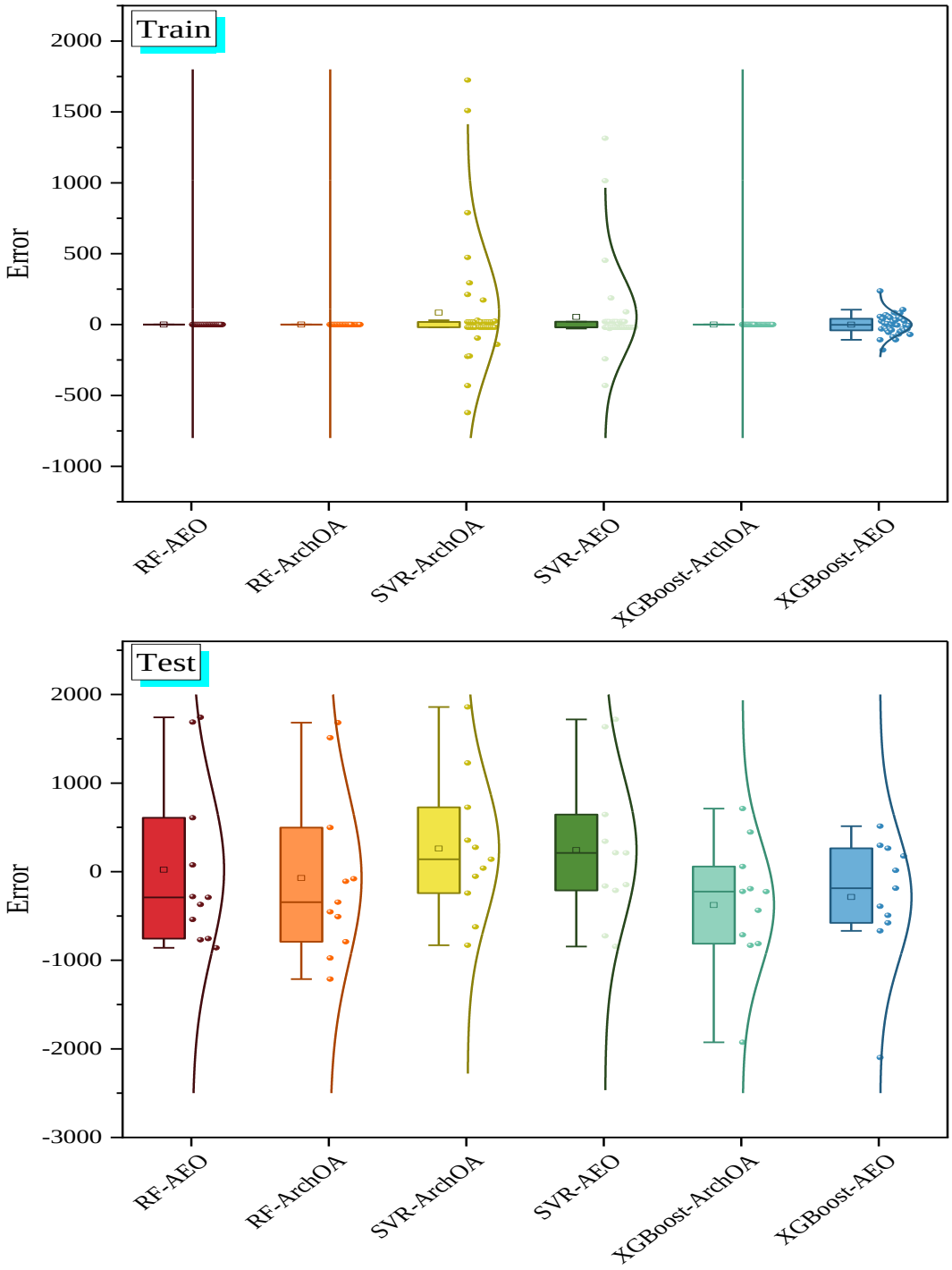

Figure 9: Box plots of error measurements for RF, SVR, and XGBoost hybrid schemes during the testing and training phases

Fig 10 depicts the temporal execution durations of each algorithm across successive iterations. The observed fluctuations in execution time suggest that algorithms with initial temporal variability tend to be less stable than those with oscillatory patterns. However, these algorithms typically converge towards more excellent stability with continued iterations. Notably, the SVR-AEO and SVR-ArchOA algorithms adhere to this trend, exhibiting shorter run times than others. This stability indicates their proximity to an optimal state in execution.

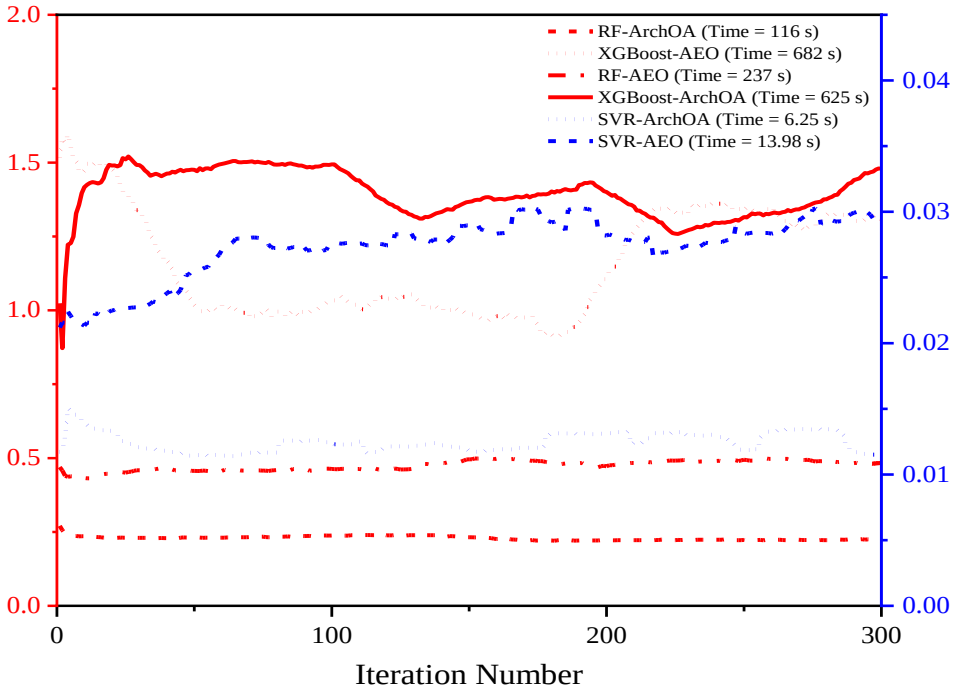

Figure 10: Comparison of runtime for various hybrid schemes

Fig 11 depicts the superior predictive model chosen for this study, evaluated based on three criteria: total run time, R^2 score, and minimum convergence, with weights of 0.5, 0.3, and 0.2. The fig demonstrates that the SVR-AEO and XGBoost-AEO schemes attain the highest levels of performance and acceptance. In contrast, the XGBoost and RF schemes exhibit comparatively lower efficiency and accuracy.

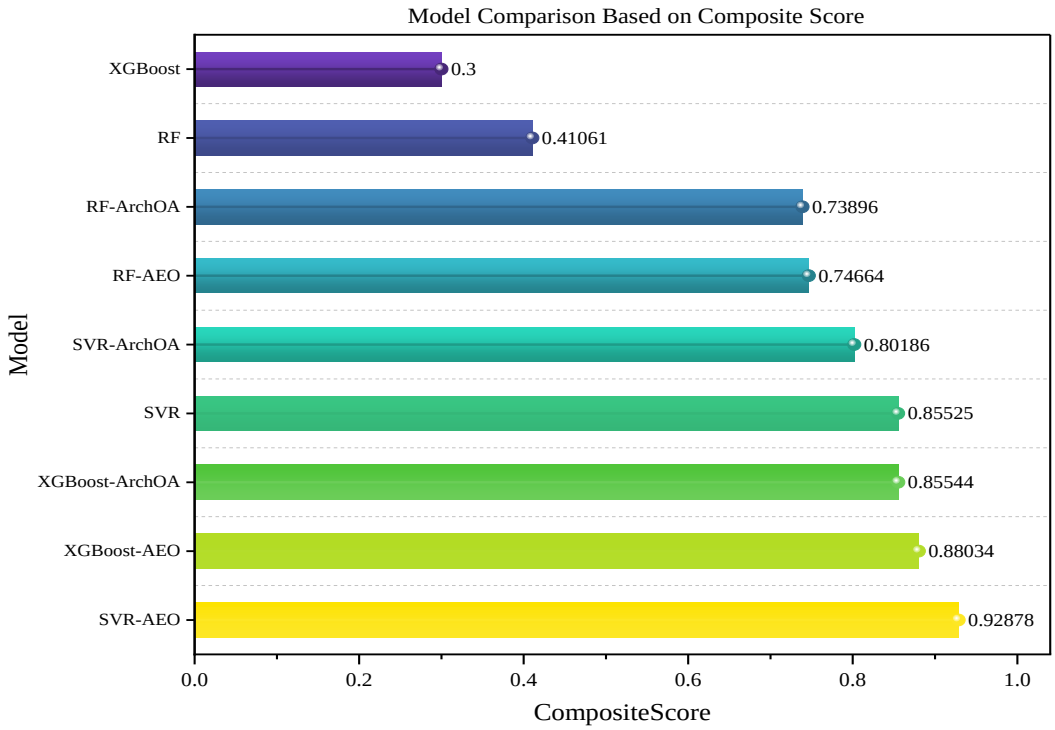

Figure 11: Model comparison based on composite score

To evaluate the influence of environmental variables on model predictions, a sensitivity analysis was conducted using the Delta Moment-Independent Measure, as shown in Table 7. This table presents the computed delta and S1 sensitivity indices along with their confidence intervals (delta_conf and S1_conf) for each input feature. Results indicate that temperature-related variables, particularly Cumu_Temp_win and Cumu_Temp_year, exhibit the highest delta values (0.36 and 0.24), suggesting strong global influence on yield predictions. In contrast, features

like Rainfall_sum and Day_over_37°C showed minimal impact (delta < 0.03), indicating limited predictive contribution. Complementing the table 6, Figure 12 visualizes these sensitivity indices using bar plots with

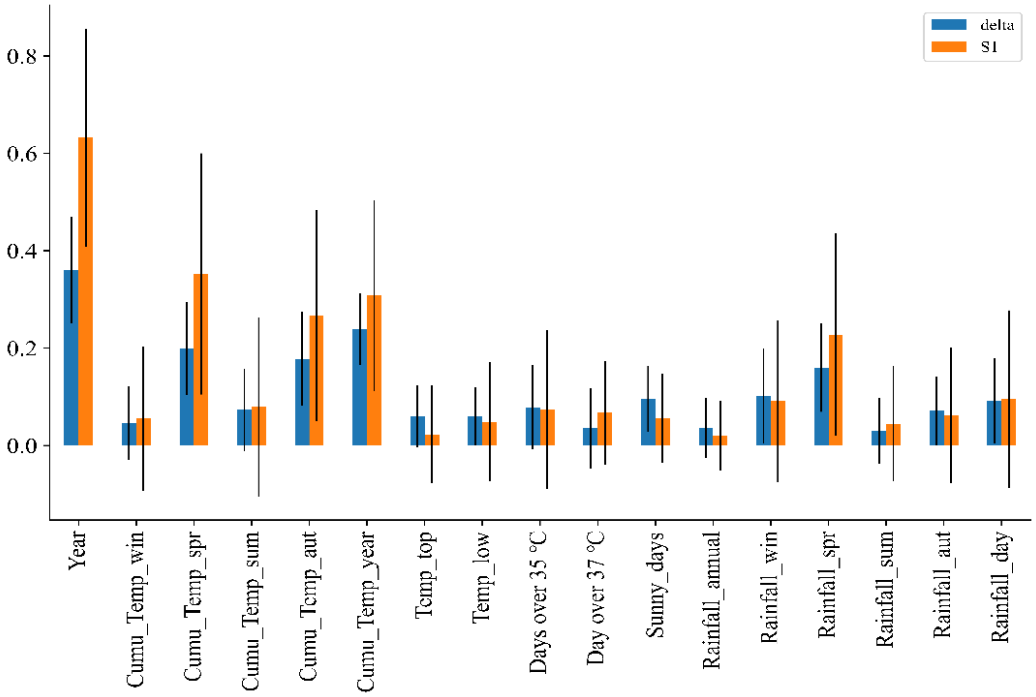

Figure 12: Features the Delta sensitivity analysis result

error bars representing uncertainty bounds. It clearly demonstrates that temperature variables dominate both delta and S1 measures, followed by moderate contributions from rainfall variables such as Rainfall_spr and Rainfall_day. The robustness of this analysis lies in its model-agnostic nature, ability to capture non-linear interactions, and provision of quantifiable uncertainty, offering a comprehensive understanding of the input-output relationship in crop yield forecasting.

Table 7: Delta and S1 sensitivity measures for input features

	delta	delta_conf	S1	S1_conf
Year	0/3602979	0/108672505	0/632609896	0/223767695
Cumu-Temp-win	0/046077548	0/074770124	0/055385049	0/14825081
Cumu-Temp-spr	0/199149499	0/094654481	0/352355572	0/247731198
Cumu-Temp-sum	0/072934084	0/084748716	0/079012135	0/184106117
Cumu-Temp-aut	0/177999487	0/09628438	0/266851715	0/21658509
Cumu-Temp-year	0/238762036	0/073477282	0/307604744	0/195738844
Temp-top	0/060479488	0/063659381	0/022610382	0/10062539
Temp-low	0/05994314	0/059953434	0/048870158	0/122519414
Days over 35 °C	0/078556725	0/08642804	0/07339147	0/163629978
Day over 37 °C	0/035495826	0/082978932	0/067183232	0/106573313
Sunny-days	0/095759529	0/067027385	0/055126859	0/091403379
Rainfall-annual	0/036150721	0/062091646	0/020485492	0/071579759
Rainfall-win	0/102089888	0/097648603	0/090905634	0/166548971
Rainfall-spr	0/159422726	0/090562472	0/227555234	0/208187665
Rainfall-sum	0/029803097	0/067603127	0/044162581	0/118354992
Rainfall-aut	0/071208772	0/070810694	0/062254171	0/139367323
Rainfall-day	0/091945193	0/087927413	0/095401637	0/182325703

4 Discussion

In this section, outlines the rationale behind the selection of the applied algorithms, highlighting their suitability for structured agricultural data. The algorithms selected in this study were chosen for their complementary strengths and suitability for structured agricultural data. RF offers robustness to noise and clear interpretability through feature importance. SVR generalizes well in small, high-dimensional datasets using kernel methods, while XGBoost provides high accuracy with built-in regularization, effectively modeling complex interactions such as temperature effects. For optimization, AEO was used for global hyperparameter search, and ArchOA for precise local refinement. Combining these models with their respective optimizers enables a balance between predictive accuracy, computational efficiency, and scalability—key requirements for reliable hickory yield forecasting. Subsequently, it presents a comparative analysis of the proposed hybrid models' performance against state-of-the-art approaches. Specifically, the evaluation includes baseline models such as XGBoost, Random Forest (RF), and Support Vector Regression (SVR), as well as hybrid models built upon these methods. the experiments show that the proposed hybrid models, particularly the XGBoost-AEO, outperform the traditional models in terms of prediction accuracy, as evidenced by the lower RMSE and higher R^2 values observed. Specifically, the XGBoost-AEO model achieved an RMSE of 742.607 and R^2 of 0.959451 on the test dataset, which is significantly better than the baseline models such as RF (RMSE: 2070.594, R^2: 0.68475) and SVR (RMSE: 1174.096, R^2: 0.898639). The observed performance improvements can be attributed to several factors. First, the hybridization of XGBoost with advanced optimization techniques (AEO) enhances its ability to capture complex patterns in the data, which simpler models like RF and SVR fail to do effectively. Additionally, the dataset characteristics, such as the number of samples and the variability in environmental conditions, play a crucial role in the performance of these models. To evaluate the real-world feasibility of deploying the proposed hybrid models, a detailed computational complexity analysis was conducted. All experiments were performed on a system with the following hardware specifications: 12th Gen Intel® Core™ i5-12400F @ 2.50 GHz, 32 GB DDR4 RAM, running Windows 10 Pro (64-bit, Build 19045.2311). This ensures consistency and reliability in execution time comparisons across models. The computational efficiency of the models, as reported in Table 6 and visualized in Figure 10, revealed substantial differences in training time, reflecting the inherent complexity of each model. XGBoost-AEO, while achieving the highest predictive performance ($R^2 = 0.959$), required approximately 682 seconds to train. This extended runtime is attributed to its ensemble-based structure, which involves extensive tree construction and hyperparameter tuning. In contrast, SVR-AEO demonstrated remarkable computational efficiency, completing its training in just 13.98 seconds, making it approximately 53 times faster than XGBoost-AEO and 15 times faster than RF-AEO, while still achieving competitive accuracy ($R^2 = 0.949$). RF-AEO offered a balanced compromise, delivering solid predictive accuracy ($R^2 = 0.941$) with a moderate execution time of 237 seconds. These findings underscore a clear trade-off between accuracy and computational cost. For precision-critical applications, such as high-stakes yield forecasting in large-scale agricultural planning, the computational expense of XGBoost-AEO is justified. However, for real-time or resource-constrained environments, SVR-AEO offers a highly efficient and accurate alternative. RF-AEO, sitting between these two extremes, provides a robust and versatile option, particularly suitable for mid-scale datasets or mixed-priority tasks where both speed and accuracy are important. Overall, these observations highlight the necessity of aligning model selection with the computational budget, application scale, and required predictive precision in agricultural forecasting systems. Hybrid models, by integrating multiple methods, can exploit the strengths of each, leading to more accurate predictions. In terms of computational efficiency, the hybrid models do introduce additional overhead compared to standalone machine learning models. However, this overhead is justified by the significant improvement in predictive performance. A detailed computational complexity analysis was conducted to evaluate execution time differences across models and to assess the trade-offs between computational cost and predictive accuracy. The findings reveal that while XGBoost-AEO achieves the highest predictive performance, it requires a considerably longer training time (842 seconds) due to its ensemble structure and complex hyperparameter tuning. In contrast, SVR-AEO demonstrated competitive accuracy with significantly lower computational cost (15.98 seconds), making it suitable for real-time or resource-constrained applications. RF-AEO offered a balanced profile, combining robust accuracy with moderate training time (237 seconds), making it a practical choice for mid-scale prediction tasks. These results highlight the importance of selecting models not only based on accuracy but also on their computational feasibility, particularly in operational environments where speed and scalability are critical. the complexity of hybrid models increases the time required for training and inference, but the gains in accuracy make it a worthwhile trade-off in applications where high prediction precision is critical. Further optimization techniques could help reduce the computational burden of these models without sacrificing accuracy. Another issue under discussion is common to observe a performance gap between the training and testing phases in machine learning models. For some models, such as Random Forest (RF) and Support Vector Regression (SVR), there is a noticeable drop in performance when tested on unseen data compared to training data. This discrepancy can often be attributed to overfitting, where the model learns noise and specific patterns from the training set that do not generalize well to new data. In contrast, models like XGBoost-AEO demonstrate a smaller gap between training and testing performance, indicating better

generalization. This suggests that the hybrid optimization techniques used in XGBoost-AEO help improve the model's ability to perform reliably on new, unseen data.

5 Conclusion

The current study identifies methods for predicting hickory yield by evaluating single schemes- random forest, support vector regression, and XGBoost- and optimizing hybrid schemes using artificial ecosystem-based optimization and the Archimedes optimization algorithm. The results are visualized through charts and tables that pinpoint key drivers of yield impact, such as temperature and rainfall, to enhance the accuracy of the predictions. AEO significantly improved the SVR and XGBoost model, with the best results and an initial R^2 of 0.6331, improving to 0. 9594 post-optimizations. The R^2 value improvement with the XGBoost-AEO hybrid model is a significant achievement in agricultural forecasting. This improvement means the model can explain more variability in crop yield data, enabling better decision-making for farmers and policymakers. A higher R^2 enhances the ability to predict yields more accurately across varying environmental conditions, which is crucial for optimizing agricultural practices and resource management. Hybrid XGBoost schemes also show enhanced predictive accuracy, notably reducing error rates. Evaluation metrics, including R^2 and MBE, consistently highlight the SVR-AEO and XGBoost-AEO model's efficacy across statistical indices, supported by stability in temporal execution analyses. Moreover, the robustness of the models across different environmental conditions is essential for ensuring their generalizability and real-world applicability. In this study, the hybrid models were tested on data that represented a range of climatic factors, and the results suggest that the XGBoost-AEO model performs consistently well across different environments. One potential limitation of the proposed models is their generalizability to other crops and regions. While the hybrid models demonstrated strong performance in predicting the crop yield for the specific case studied, their effectiveness may vary in other agricultural contexts. Factors such as soil type, climatic conditions, and regional farming practices could influence the model's performance. However, in future works further testing on additional datasets from diverse geographic regions and climate zones would be beneficial to fully evaluate the model's robustness. Overall, the study underscores the significance of model selection and optimization in improving forecasting precision, which is crucial for advancing agricultural sustainability and productivity in hickory cultivation.

Nomenclature

ANNs	Artificial Neural Networks	R2	Coefficient of Determination
CART	Classification and Regression Trees	RF	Random Forest
DF	best fitness	SVR	Support Vector Regressions
EFA	Exploratory factor analysis	tmax	maximum number of iterations
F	factors	VAF	Variance Accounted For
$f_k(x_i)$	score assigned	W	Weights for each factor
GBM	Gradient Boosting Machines	X	input vector
MAE	Mean Absolute Error	XGBoost	Extreme Gradient Boosting
K	number of factors	M	schemes
MBE	Mean Bias Error		
ML	Machine Learning		
MSE	Mean Square Error		
NRMSE	Normalized Root Mean Square Error		

Competing of interests

The authors declare no competing of interests.

Authorship contribution statement

Xu Guo: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Data availability

On Request

Declarations

Not applicable

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Funding

Not applicable

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

References

[1] Phalan, B., R. Green and A. Balmford (2014). Closing yield gaps: perils and possibilities for biodiversity conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, The Royal Society Publishing, 369(1639), p. 20120285. https://doi.org/10.1098/rstb.2012.0285.

[2] Tilman, D., C. Balzer, J. Hill and B.L. Befort (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, National Academy of Sciences (NAS), 108(50), pp. 20260–20264. https://doi.org/10.1073/pnas.1116437108.

[3] Basso, B. and L. Liu (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, Elsevier, 154, pp. 201–255. https://doi.org/10.1016/bs.agron.2018.11.002.

[4] Chipanshi, A., Y. Zhang, L. Kouadio, N. Newlands, A. Davidson, H. Hill, R. Warren, B. Qian, B. Daneshfar and F. Bedard (2015). Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. *Agricultural and Forest Meteorology*, Elsevier, 206, pp. 137–150. https://doi.org/10.1016/j.agrformet.2015.03.007.

[5] García-León, D., R. López-Lozano, A. Toreti and M. Zampieri (2020). Local-scale cereal yield forecasting in Italy: Lessons from different statistical models and spatial aggregations. *Agronomy*, MDPI, 10(6), pp. 809. https://doi.org/10.3390/agronomy10060809.

[6] Mishra, S., D. Mishra, P.K. Mallick, G.H. Santra and S. Kumar (2021). A novel borda count based feature ranking and feature fusion strategy to attain effective climatic features for rice yield prediction. *Informatica*, Slovenian Society Informatika, 45(1). https://doi.org/10.31449/inf.v45i1.3258.

[7] Becker-Reshef, I., C. Justice, B. Barker, M. Humber, F. Rembold, R. Bonifacio, M. Zappacosta, M. Budde, T. Magadzire and C. Shitote (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. *Remote Sensing of Environment*, Elsevier, 237, p. 111553. https://doi.org/10.1016/j.rse.2019.111553.

[8] Fao, F (2018). *Food and agriculture organization of the United Nations. Rome*. Http://Faostat. Fao. Org, 403

[9] Schauberger, B., J. Jägermeyr and C. Gornott (2020). A systematic review of local to regional yield forecasting approaches and frequently used data resources. *European Journal of Agronomy*, Elsevier, 120, p. 126153. https://doi.org/10.1016/j.eja.2020.126153.

[10] Tubiello, F.N., M. Salvatore, R.D. Cóndor Golec, A. Ferrara, S. Rossi, R. Biancalani, S. Federici, H. Jacobs and A. Flammini (2014). Agriculture, forestry and other land use emissions by sources and removals by sinks. Rome, Italy. https://catalogue.unccd.int/356_i3671e.pdf.

[11] Rembold, F., C. Atzberger, I. Savin and O. Rojas (2013). Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing*, MDPI, 5(4), pp. 1704–1733. https://doi.org/10.3390/rs5041704.

[12] Cai, Y., K. Guan, D. Lobell, A.B. Potgieter, S. Wang, J. Peng, T. Xu, S. Asseng, Y. Zhang and L. You (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, Elsevier, 274, pp. 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

[13] Johnson, M.D., W.W. Hsieh, A.J. Cannon, A. Davidson and F. Bédard (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, Elsevier, 218, pp. 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003

[14] Kamir, E., F. Waldner and Z. Hochman (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, 160, pp. 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008.

[15] Mateo-Sanchis, A., M. Piles, J. Muñoz-Marí, J.E. Adsuara, A. Pérez-Suay and G. Camps-Valls (2019). Synergistic integration of optical and microwave satellite data for crop yield estimation. *Remote Sensing of Environment*, Elsevier, 234, pp. 111460. https://doi.org/10.1016/j.rse.2019.111460.

[16] Wolanin, A., G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi and L. Guanter (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, IOP Science, 15(2), pp. 24019. DOI: 10.1088/1748-9326/ab68ac

[17] Zhang, L., Z. Zhang, Y. Luo, J. Cao and F. Tao (2019). Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning approaches. *Remote Sensing*, MDPI, 12(1), pp. 21. https://doi.org/10.3390/rs12010021.

[18] Goodfellow, I., Y. Bengio and A. Courville (2016). *Deep learning*. MIT press. https://doi.org/10.4258/hir.2016.22.4.351.

[19] Liu, H., Original data for train and test.

[20] Breiman, L (2001). Random forests. *Machine Learning*, Springer Nature, 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324.

[21] Archer, K.J. and R. V Kimes (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, Elsevier, 52(4), pp. 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

[22] Horning, N (2010). Random Forests: An algorithm for image classification and generation of continuous fields data sets, In *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences*, Osaka, Japan, pp. 1–6.

[23] Liaw, A. and M. Wiener (2002). Classification and regression by randomForest. *R News*, 2(3), pp. 18–22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf.

[24] Pal, M (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, Taylor & Francis, 26(1), pp. 217–222. https://doi.org/10.1080/01431160412331269698.

[25] Watts, J.D., S.L. Powell, R.L. Lawrence and T. Hilker (2011). Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. *Remote Sensing of Environment*, Elsevier, 115(1), pp. 66–75. https://doi.org/10.1016/j.rse.2010.08.005.

[26] Liaw, A (2002). Wiener M. Classification and Regression by Randomforest R News, 2(3), pp. 18. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf.

[27] Sabzekar, M. and S.M.H. Hasheminejad (2021). Robust regression using support vector regressions. *Chaos, Solitons and Fractals*, Elsevier, 144, pp. 110738. https://doi.org/10.1016/j.chaos.2021.110738.

[28] Bisri, A. and R.S. Wahono (2015). Penerapan Adaboost untuk penyelesaian ketidakseimbangan kelas pada Penentuan kelulusan mahasiswa dengan metode Decision Tree. *Journal of Intelligent Systems*, 1(1), pp. 27–32. https://www.neliti.com/publications/243690/penerapan-adaboost-untuk-penyelesaian-ketidakseimbangan-kelas-pada-penentuan-kel.

[29] Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system, In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, ACM SIGKDD, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[30] Zhao, W., L. Wang and Z. Zhang (2020). Artificial ecosystem-based optimization: a novel nature-inspired meta-heuristic algorithm. *Neural Computing and Applications*, Springer Nature, 32(13), pp. 9383–9425. https://doi.org/10.1007/s00521-019-04452-x.

[31] Hashim, F.A., K. Hussain, E.H. Houssein, M.S. Mabrouk and W. Al-Atabany (2021). Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. *Applied Intelligence*, Springer Nature, 51, pp. 1531–1551. https://doi.org/10.1007/s10489-020-01893-z.

[32] BANA E COSTA, C.A. and J. VANSNICK (1997). Applications of the MACBETH approach in the framework of an additive aggregation model. *Journal of Multi-Criteria Decision Analysis*, WILEY Online Library, 6(2), pp. 107–114. https://doi.org/10.1002/(SICI)1099-1360(199703)6:2%3C107::AID-MCDA147%3E3.0.CO;2-1.

[33] Krajewski, L.J. and L.P. Ritzman (1999). *Operations management: strategy and analysis*. Addison-Wesley New York. https://library.wur.nl/WebQuery/titel/1617981.

[34] Barron, F.H. and B.E. Barrett (1996). Decision quality using ranked attribute weights. *Management Science*, Informs PubsOnline, 42(11), pp. 1515–1523. https://doi.org/10.1287/mnsc.42.11.1515.

[35] Olson, D.L (1997). Decision aids for selection problems. *Journal of the Operational Research Society*, Taylor & Francis, 48(5), pp. 541–542. https://doi.org/10.1057/palgrave.jors.2600636.

[36] Fabrigar, L.R. and D.T. Wegener (2011). *Exploratory factor analysis*. Oxford University Press.

[37] Rastgoo, A. and H. Khajavi (2023). A novel study on forecasting the Airfoil self-noise, using a hybrid model based on the combination of CatBoost and Arithmetic Optimization Algorithm. *Expert Systems with Applications*, Elsevier, p. 120576. https://doi.org/10.1016/j.eswa.2023.120576.