Organic Farming
https://www.acadlore.com/journals/OF

Forecasting Yield of Coffee Crop Varieties C×R, Sln3 and Sln5B: A Stochastic Machine Learning Model Based on Agro-Ecological Factors using Multivariate Feature Selection Approach

Chandagalu Shivalingaiah Santhosh[1*], Kattekyathanahalli Kalegowda Umesh[2], Venkatesh Hemanth[1], Khatri Narendra[3]

[1] Department of Computer Applications, JSS Science and Technology University, 570006 Mysuru, India
[2] Department of Information Science & Engineering, JSS Science and Technology University, 570006 Mysuru, India
[3] Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of Higher Education, 576104 Manipal, India

[*] Correspondence: Chandagalu Shivalingaiah Santhosh (sancs84@jssstuniv.in)

Received: 06-18-2025 **Revised:** 09-03-2025 **Accepted:** 09-10-2025

Citation: Santhosh, C. S., Umesh, K. K., Hemanth, V., & Narendra, K. (2025). Forecasting yield of coffee crop varieties C×R, Sln3 and Sln5B: A stochastic machine learning model based on agro-ecological factors using multivariate feature selection approach. *Org. Farming*, *11*(3), 203–226. https://doi.org/10.56578/of110305.

Abstract: Accurate forecasting of coffee crop yield is essential for enhancing agricultural decision-making, ensuring food security, and mitigating environmental risks. India cultivates both Arabica and Robusta across more than one hundred registered varieties. In this study, yield forecasts were developed for three representative varieties—C×R, Sln3, and Sln5B—using agro-ecological data collected from 2015 to 2022 at the Central Coffee Research Institute (CCRI), Coffee Research Station, Balehonnur, Karnataka, India. A stochastic machine learning framework was employed to identify and evaluate the most influential agro-ecological predictors through a multivariate feature selection approach coupled with correlation matrix analysis. Optimal predictors were organized into three distinct parameter groups, which were then used as inputs to four regression models: Extra Trees (ET), Gradient Boosting (GB), Random Forest (RF), and Decision Tree (DT). Independent testing revealed that the ET model consistently provided the highest accuracy. For C×R, yield was most accurately predicted using Group-1 parameters, such as coffee leaf rust (CLR), minimum temperature (Tmin), maximum temperature (Tmax), relative humidity (Rh), rainfall (Rf), organic carbon (OC), phosphorus (P), potassium (K), pH, plant spacing (Sp), and plant age (Ag), achieving a coefficient of determination (R^2) of 0.98 with a Root Mean Square Error (RMSE) of 8.61 kg ha^{-1}. For Sln3, Group-3 parameters, such as CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, minimum sunshine hours (SSmin), maximum sunshine hours (SSmax), vapor (Vp), and dew point (Dp), produced an R^2 of 0.98 with an RMSE of 8.27 kg ha^{-1}, while for Sln5B, Group-3 parameters yielded an R^2 of 0.97 with an RMSE of 7.79 kg ha^{-1}. These results demonstrate the superiority of the ET algorithm compared with GB, RF, and DT models, which exhibited comparatively lower predictive accuracy. Simulation outcomes further revealed that age, rainfall, and the incidence of CLR were among the most decisive agro-ecological determinants of yield. These findings underscore the potential of stochastic machine learning models, particularly the ET model, for enhancing yield prediction and identifying agro-ecological drivers of coffee productivity.

Keywords: Coffee; Stochastic approach; Machine learning; Yield prediction; Agro-ecological factors

1. Introduction

Coffee is a major export commodity and foreign currency generator. It is a growing export crop in India. Coffee's social, institutional, and cultural roles are extremely important in rural southern India. India has 103 coffee bean species, although Arabica and Robusta are the most significant. For coffee, India ranks seventh in area, sixth in output, third in productivity, and sixth in export. Santhosh & Umesh (2022) forecasted coffee production using 2008–2020 Chikkamagaluru agronomic soil fertility data. Using multiple linear regression (MLR), Lasso regression, and elastic net regression, this research predicted coffee yield. Model output was evaluated using

coefficient of determination (R^2) and Root Mean Square Error (RMSE). Elastic net regression predicts coffee output better than Lasso and MLR models, and agro-ecological conditions in Chikkamagaluru are expected to diminish productivity by 15–25% by 2050. Maro et al. (2014) conducted a study in Tanzania to address the significant coffee yield losses caused by poor soil conditions. A straightforward, quantitative technique was employed to analyze coffee productivity and to provide nutrient management recommendations. The research was conducted by the Tanzania Coffee Research Institute (TACRI), Lyamungu, using 2010–2013 data from Hai and Lushoto in Northern Tanzania. The new method, Soil Analysis for Fertility Evaluation and Recommendation on Nutrient Application to Coffee (SAFERNAC), was developed. The model comprises SOIL (basic soil qualities), PLANT (plants and corresponding management parameters, such as fertilizer absorption, density, and maximum returns per tree), and INPUT. The model achieved an accuracy of 80–100% when tested against adjusted equations.

Kouadio et al. (2018) analyzed clay richness and coffee yield using extreme learning machine feed-forward network models. 18 extreme learning machine models were tested using various soil conditions. Finally, the output of those models was compared with that of Random Forest (RF) and MLR using R^2 and RMSE values. Santhosh & Umesh (2023) proposed machine learning-based ensemble models to predict Chikkamagaluru coffee harvests. Yield and soil factor data from 2008 to 2022 were used to forecast the production. Using accuracy and Receiver Operating Characteristic (ROC) curve values, the ensemble models were evaluated, including Gradient Boosting (GB), Extra Trees (ET) and RF classifiers. ET classifiers outperformed other models with an accuracy of 91%.

Romero-Alvarado et al. (2002) evaluated two shaded coffee systems for productivity and soil nutrients. Species-rich natural vegetation and Inga latibracteata Harms dominate. Producer plots were examined at the Francisco I. Madero Community, Jitotol, Chiapas. Variables measured included nitrogen (N), calcium (Ca), magnesium (Mg), organic matter, soil pH, species richness, shade-tree density, layers, tree diameter, tree height, percentage of shade cover, direct and diffused light, and coffee yield. Results indicate that further research on low-input coffee production, nutrient cycling, and N-fixation is needed to understand nutritional dynamics in the system and evaluate Inga Shade (IS) and Rustic Shade (RS) types for small farmers. Shastry et al. (2016) employed Artificial Neural Network (ANN) and MLR models to estimate wheat yield using rainfall, transpiration, biomass, extractable soil water (ESW), soil nitrogen (NO_3), soil evaporation, and historical wheat yield. The outcomes of ANN and MLR were compared using R^2 and prediction inaccuracy. MLR, Dynamic Artificial Neural Network (D-ANN), and Conventional Artificial Neural Network (C-ANN) models achieved R^2 scores of 92.52%, 95%, and 97% and average prediction errors of 4.19%, 2.24%, and 0.52% on the test set, respectively. In R^2 and percentage prediction error, C-ANN outperformed D-ANN and MLR. C-ANN predicted wheat production better than MLR and D-ANN in the data set.

Singh et al. (2017) studied machine learning models to predict yield category based on the macronutrient and micronutrient status dataset. Agricultural output projection data were obtained from Krishi Bhawan, Talab Tillo, Jammu. The dataset included micronutrients, such as zinc (Zn), iron (Fe), manganese (Mn), and copper (Cu), and macronutrients, such as pH, organic carbon (OC), electrical conductivity (EC), N, P, K, and sulfur (S), from soil samples collected in the Jammu District. Machine learning models were then applied to predict yield categories based on nutrient status. Prediction accuracy of crop production was 96.43% for the k-nearest neighbor (KNN) classifier, 97.80% for the Naïve Bayes classifier, and 93.38% for the Decision Tree (DT) classifier. Kittichotsatsawat et al. (2022) measured regions, zones, rainfall, relative humidity, lowest temperatures, and humidity to estimate yearly coffee yields and match output to market demand using ANN and MLR. $R^2 = 0.9524$ and RMSE = 0.0642 indicate an accurate prediction of cherry coffee production using MLR. Based on soil composition and climate, Aceves Navarro et al. (2018) determined Tabasco's best Robusta coffee-producing zones. The current and worst-case climate change scenarios, i.e., Representative Concentration Pathway 8.5 (RCP8.5), employed the Agro-Ecological Zoning (AEZ) of the Food and Agriculture Organization of the United Nations (FAO). Both climates' optimal yields were examined for low, medium, and high inputs. By 2050, a 1.6ºC increase in daily temperature would not affect Tabasco's ideal Robusta coffee-producing zone (RCP8.5: worst-case scenario). By 2050, higher daytime temperatures will reduce the maximum photosynthetic ratio and mean potential yields by 41%.

Campanha et al. (2004) compared agroforestry to monoculture coffee varietals after monitoring temperature, nutritional condition, coffee plant yield, and vegetative growth for two years. Area, leaves, branch length, and nodes were determined in data analysis utilizing students' statistical test at a 5% confidence level. The result showed that monoculture generated 2,443 kg ha^{-1} of coffee, whereas agroforestry produced 515 kg ha^{-1}. Agroforestry increased tree growth and lowered maximum and diurnal temperatures. Muliasari et al. (2022) estimated the Robusta coffee output in Bogor District from July to December in 2020. Using a Geographic Information System (GIS), the study analyzed the digital elevation model (DEM), agro-climate, physical and organic properties of the soil, land use, and socioeconomic factors, such as protected areas, lakes, roads, and rivers. Data processing included interpolation and classification. In Bogor District, 2% (5,227.78 ha) of land was classified as moderately appropriate (S2), 33% (99,189.20 ha) as marginal (S3), and 65% (194,808.40 ha) as unsuitable (N).

Varshitha & Choudhary (2022) used a machine learning model based on Bootstrap Aggregating (Bagging) to

predict soil fertility and crop productivity. Soil pH, temperature, humidity, OC, precipitation, N, P, K, and moisture content were examined to predict agricultural production and soil fertility. The efficacy of various models, such as Support Vector Machine (SVM), DT, Naïve Bayes, and KNN, was assessed using R^2 values from Bagging and other approaches to predict soil fertility and production. Wang et al. (2015) investigated 254 coffee plots across Uganda's Central, Northern, Eastern, Southwestern, and Northwestern coffee belts to examine key production factors and associated yield variations. Global Positioning System (GPS) was used to estimate plot sizes and positions. In addition, boundary line analysis was conducted to study coffee production characteristics and yield discrepancies. Robusta and Arabica yields in Uganda varied significantly (778, 760, 966, 994, and 774 kg ha^{-1} $year^{-1}$) compared to achievable values (1,737, 1,464, 1,701, 2,243, and 1,550 kg ha^{-1} $year^{-1}$) in five locations.

Drummond et al. (2003) studied stepwise multiple linear regression (SMLR), projection pursuit regression (PPR), and supervised feed–forward neural networks to find algorithms that could relate soil parameters and grain yields point-by-point across 10 site-years. The crops considered were corn and soy. In every site-year, SMLR and PPR were outperformed using neural methods, resulting in minor prediction errors, according to R^2. Kaul et al. (2005) studied MLR and ANN models to predict maize and soybean yields under typical meteorological conditions in Maryland. Models were developed using meteorological data from multiple locations across Maryland. The ANN model for corn yields had an R^2 and RMSE of 0.77 and 1036, compared to linear regression's 0.42% and 1.356%. In contrast to the linear regression model's R^2 of 0.46 and RMSE of 312, the ANN model for soybean yields achieved an excellent R^2 of 0.81 and RMSE of 214.

Natarajan et al. (2016) used the fuzzy cognitive map (FCM), Data Driven Nonlinear Hebbian Learning (DDNHL), and genetic algorithm to help classify sugarcane production. The FCM model was used to forecast sugarcane production in precision agriculture using soil and climatic factors. With performance metric accuracy, DDHNL had 94.7% output prediction accuracy, FCM-GA 93.4%, and FCM-DDHNL 92.1%. Those models identified sugarcane production prediction factors. Shakoor et al. (2017) developed predictive models for intelligent agriculture information systems in Bangladesh. Aman rice, Boro rice, Aus rice, potato, jute, wheat and rice were evaluated as the main crops. In the forecast, supervised machine learning was used to analyze temperature, rainfall, and yield. Iterative Dichotomiser 3 (ID3) and DT forecasted Aus rice, Aman rice, and wheat with better results. KNN performed better in predicting Boro rice, jute, and potato than ID3.

Kim et al. (2014) compared agricultural pest prediction systems based on machine learning. Bayesian, neural, MLR, and SVM algorithms and their applications were discussed. Leaf moisture, pests, and illnesses for particular crops were predicted using generalized regression neural networks, ridge regressions, lasso, MLR, Bayesian, elastic net, and Random Forest Regression (RFRs) and their outputs. Santhosh & Umesh (2024) conducted a comparative study utilizing five machine learning models incorporating biotic and abiotic variables to predict coffee crop productivity in Chikkamagaluru, Karnataka. The study evaluated the predictive performance of ET, GB, RF, DT, and KNN models. Comparison of model performance was conducted utilizing an independent testing dataset: Mean Absolute Error (MAE), Mean Square Error (MSE), RMSE, and R^2 errors. Regression models using Group 1 and 2 characteristics and fine-tuning functions accurately forecasted coffee yield (R^2 = 0.98 kg ha^{-1} and RMSE = 7.96 kg ha^{-1} and R^2 = 0.96 kg ha^{-1} and RMSE = 10.96 kg ha^{-1}). The ideal weather parameter outperformed the RF, DT, and KNN models in forecasting biotic-CLR incidence in coffee production.

Sudha et al. (2020) conducted a study at the the Central Coffee Research Institute (CCRI), Chikkamagaluru, Karnataka, India, to investigate the relationship between climatic conditions and CLR occurrence. The study utilized the leaf rust-resistant Arabica coffee selection Sln3, the interspecific hybrid Sln5B, and the Robusta cultivar C×R for comparative evaluation. In 2015–2016, 2016–2017, 2017–2018, and 2018, the CCRI farm observed CLR incidence for Coffea arabica L. cultivars Sln3 and Sln5B, and Coffea canephora cultivar C×R. The observatory collected temperature, humidity, and rainfall. Regression analysis of climate and leaf rust incidence was conducted using MS Excel. Cerda et al. (2017) quantified primary and secondary yield losses in coffee caused by pests and diseases and identified the major contributing factors. Researchers constructed a full-sun coffee packaging with six pesticide spray regimes. Pests and diseases reduced primary production by 26% and secondary output by 38%.

In conclusion, researchers have advised tying expected yield losses to production variables to help farmers reduce losses. Tadesse et al. (2021) reported that CLR is the most devastating disease affecting coffee worldwide. Climate change directly threatens agricultural productivity via plant growth and production changes and indirectly affects crop diseases. Climate change has cut global agricultural production by 1% to 5% every decade for 30 years. The main meteorological parameters affecting CLR are wind speed, humidity, and temperature. Higher average low temperatures tend to make more locations at higher elevations suitable for CLR, which increases outbreaks and severity.

Despite the global significance of coffee, research on yield forecasting has largely focused on countries such as Tanzania, Uganda, Mexico, and Vietnam, with limited emphasis on India. The few Indian studies have examined Chikkamagaluru datasets using partial agro-ecological factors, such as soil fertility or climatic variables, often employing traditional regression or basic machine learning techniques. The integration of soil, climate, and management factors into predictive models remains underexplored. Moreover, advanced stochastic and ensemble-

based artificial intelligence (AI) approaches, capable of capturing nonlinear and complex interactions influencing coffee yields, have rarely been applied in the Indian context. The present study aims to address these gaps by applying stochastic AI models—RFR, Extra Trees Regression (ETR), Gradient Boosting Regression (GBR), and Decision Tree Regression (DTR)—to forecast coffee yields using comprehensive agro-ecological data from the CCRI, Balehonnur, Karnataka. By integrating soil, climatic, and agronomic parameters into a unified framework, this study demonstrates the potential of AI-driven models to outperform conventional approaches. The objective is to provide a robust decision-support tool for coffee growers, enabling improved yield prediction and management under variable agro-ecological and climate change conditions.

2. Materials and Methods

2.1 Area of Study and Dataset

Data from the period 2015–2016 to 2022–2023 were collected from the CCRI in Balehonnur, Karnataka, India. The dataset for the model employed in the current investigation consists of agro-ecological parameters, totaling 15 input characteristics of coffee-growing blocks in the coffee research station, Balehonnur. Table 1 depicts the dataset and provides a detailed description of the 15 agro-ecological and yield parameters. The dataset was prepared by manually collecting recorded ledger entry records at the CCRI, the coffee research station at Balehonnur, Chikamagaluru, and the domain expert's inputs. The relationships among the various factors influencing coffee crop yield were analyzed. A total of 576 samples were collected for the research, which includes three coffee types: Sln3, Sln5B, and C×R CLR incidence data, which is of 192 samples each (192 samples × 3 coffee types = 576 samples). Later, for implementation, these samples were grouped into Group-1, Group-2, and Group-3 as input parameters for models to perform yield prediction based on a multivariate feature selection approach, along with a correlation matrix output and a feature importance function considered during model implementation, which is explained in Table 1.

Table 1. Dataset of the agro-ecological factors

Dataset Categories	Model Subscript Abbreviation	Factor Name	Descriptions of Critical Intervals in the Datasets
Agronomic factors	Yf	Year from	2015
	Yt	Year to	2022
	M	Month	January to December
	Ag	Age (yr old)	4–85
	Sp	Spacing (m)	1.5–3.5
	OC	Organic carbon (kg ha^{-1})	2.7–7.8
	K	Potassium (kg ha^{-1})	63–930
	P	Phosphorus (kg ha^{-1})	9–114
	pH	Acidity level	4.5–7
Abiotic factors	Tmin	Minimum temperature (°C)	12.4–22.8
	Tmax	Maximum temperature (°C)	20.6–29.8
	SSmin	Minimum sunshine hours (h)	1–6
	SSmax	Maximum sunshine hours (h)	4.5–8
	Rf	Rainfall (cm)	186.1–329.5
	Rh	Relative humidity (%)	83–100
	Vp	Vapor (%)	16.7–25.63
	Dp	Dew point (°C)	14.8–21.0
Biotic factors	CLR incidence	CLR incidence in three coffee varieties: Sln3 & Sln5B (Arabica) and C×R (Robusta)	Incidence data were recorded twice per month, corresponding to two fortnights, and expressed as percentages (0–79.69%).
	Fn	• Day 1 to Day 15: First fortnight. • Day 16 to the last day of the respective month: Second fortnight.	
Respective yield of CCRI considered (kg ha^{-1})	Y	210–400	

2.2 Proposed Methodology

Stochastic machine learning techniques model and forecast coffee output based on agro-ecological parameters. The ideal procedure is illustrated in Figure 1. The following procedures were used for pre-processing of the proposed method. Missing data in the initially collected samples were identified. Statistical techniques such as

binning, arithmetic median, and average were used to fill in blanks where data were unavailable. Based on the standard critical values for agronomic, abiotic, and biotic parameters supplied by domain specialists at CCRI, the dataset is considered to be normalized.

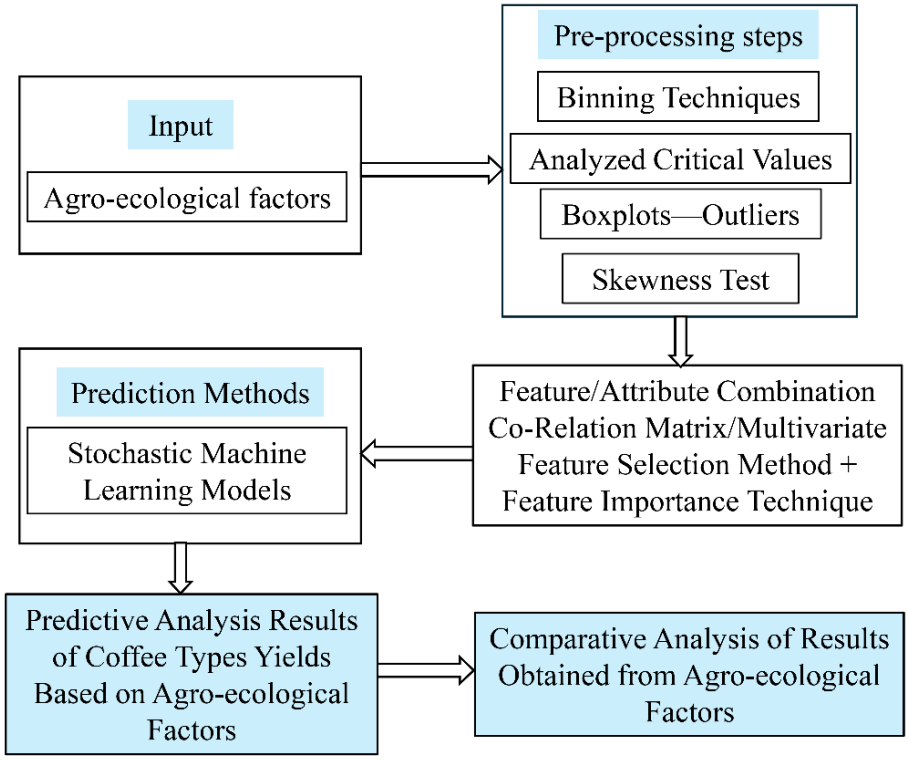

Figure 1. Block diagram of the proposed approach for predicting coffee yield based on agro-ecological factors

Applying Pearson's coefficient function and producing a correlation matrix with the Python library help to check the features positively correlated with yield. The most affecting features/predictable variables for yield were determined using a multivariate feature selection technique, which selects the best features based on statistical tests. For the pre-processing of an estimator, SelectKBest was used to eliminate all except the *K* highest-scoring characteristics for prediction in univariate analysis. In addition, box plots were used to express input features and remove outliers. A normal distribution, unaffected by skewness or outliers, is assured by the normality test. In continuation, a correlation matrix was generated to understand the correlations between input variables and the importance of features for coffee yield prediction.

To pick the most prominent characteristics and increase yield, 32,768 input feature combinations were produced and evaluated using univariate and multivariate feature selection techniques. After Python's feature significance function and correlation matrix output analysis, three best feature groups were chosen, which were grouped into Group-1, Group-2, and Group-3 input parameters and utilized to create the prediction model to assess yield output. The correlation matrix depicts the degree and direction of relationships between numerous items as coefficients from –1 (strongly negative) to +1 (strongly positive), with zero indicating no link. After the feature selection step, the grouped input parameters were given as inputs to the proposed stochastic machine learning models to predict the actual yield based on historic data. Based on the performance metrics such as R^2, MAE, MSE and RMSE, models were evaluated considering three coffee varieties. Comparative analysis was performed considering proposed stochastic machine learning models through scatter plot visualization based on R^2 and RMSE values.

2.3 Methods

To evaluate and forecast the coffee crop output of the Chikkamagaluru area, Karnataka, by taking into account the agro-ecological parameters of each region, this empirical research suggested a set of regression models based on stochastic processes. This study used four distinct stochastic tree-based regression models to examine the fundamental connections between the various agro-ecological elements. Therefore, it was possible to predict the coffee crop production using the data generated by these stochastic models for each independent variable. The models' performance was compared using several metrics, including R^2, MAE, MSE, and RMSE. Regarding coffee yield prediction, the regression-based stochastic models achieved performance comparable to state-of-the-

art models in coffee yield prediction.

2.3.1 ETR

A novel RF model employs ETR in this case. ETR generates regression trees or unpruned conclusions. The RFR method makes use of bagging and bootstrapping (Geurts et al., 2006). The instructions below show how to use the ETR algorithm for numerical characteristics:

Step 1: Dividing a node (*A*)

- The input is the learning subset *A* for the neighboring node that has to be divided.
- A node split [*x xy*] or a zero split is the output.
- Return zero (0) if the stop split (*A*) equals.
- Anyhow, from all non-consistent (in *A*) applicant characteristics, choose *B* attributes (c_1... c_b).
- Describe the locations of the *B* divides, e_1...e_b. e_i, by choosing a random number.
- Split (*A*, x_i), with v_j = 1, and then *B*;
- To ensure that *Count* (*d**, *A*) = max_i = 1... *B Count* (*ei*, *A*), return a split *e**.

Step 2: Take a haphazard split (*A*, *x*)

- Inputs include an attribute *x* and an *x* subclass *A*.
- Results: *x* split.
- Let *xAmax* and *xAmin* denote the maximum and minimum estimates of *m* in *A*, respectively.
- Illustrate any boundary *ac* in accordance with [*xAmin*, *xAmax*].
- Send the split ([*x xy*]) back.

Step 3: **Reverse split (*A*)**

- Enter: *x* subclass *A*, binary output for *x*; return TRUE if |*A*| *xmin*; return TRUE if *A*'s properties are all consistent.
- Return TRUE if the result is in accordance with *A*; FALSE otherwise.

The steps above describe the ET splitting approach for numerical features.

2.3.2 GBR

In ensemble learning, GBR tree approaches employ weak learner regression trees (DTs) to create reliable forecasting models. Poorly trained models (regressors or classifiers) have fewer errors (Singh et al., 2021). The GB tree, known as the $F_n(x_t)$ algorithm, accumulates *n* regression trees:

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \tag{1}$$

Every function $f_i(x_t)$ is a DT or regression. The equation estimates the new DT $f_{n+1}(x_t)$ to form the ensemble of trees:

$$argmin \sum_{t} L(y_t \cdot F_{\mathrm{n}}(x_t) + f_{n+1}(x_t)) \tag{2}$$

where, the loss function *L* (.) is differentiable. The steepest descent method solves this optimization. This study employed a 0.2 learning rate and a 100 estimator value. When the learning rate is smaller, it is easier to stop before over-fitting.

2.3.3 RFR

Ensemble-based data mining, like RF, makes accurate predictions without over-fitting. They use model aggregation-based learning (Kouadio et al., 2018). RFs combine binary DTs, bootstrapped learning samples, and a random explanatory variable collection. The RF method builds up to 2,000 random trees using validation or "out-of-bag" predictions in a third of samples. Each tree is bootstrapped (Kouadio et al., 2018). The RF algorithm is carried out sequentially as follows:

Step 1: Randomly choose and replace a selection of *N* samples from the training set. The training process involves growing the original trees.

Step 2: RF randomly selects *m* variables from each node's inputs (or predictors).

Step 3: The RF approach grows each tree to its greatest size without trimming its structure.

Step 4: Pooling *n* trees' predictions provides a mean value for forecasting new data if there is a regression issue.

2.3.4 DTR

Prediction trees use known inputs and feature values to calculate response or class time instants ahead. A binary tree creates hierarchical decision-making by applying tests to inputs at each node, depending on the results (Chowdhury et al., 2018). There are three stages to creating a tree by hierarchically partitioning the space: dividing nodes, identifying terminal nodes, and labeling or anticipating values at terminal nodes. Some class labels or

predicted values are more likely or expensive than others, depending on the majority or weighted votes. To make a forecast, the structure undergoes a recursive partitioning or sub-division procedure that uses binary questions for each feature value to build the tree's future branches.

2.4 Predictive Model Development

All four models—DTR, ETR, GBR, and RFR—were created on Windows 11 using Collab and Jupyter on an Intel® Core i7 laptop. To predict C×R, Sln3 and Sln5B yield (*Y*) in CCRI, the coffee research station at Chikkamagaluru, the DTR model used patterns embedded in the data matrix of *K* (= 15) lots of agro-ecological factors, which combines agronomic, abiotic, and biotic factors from Table 1 and its relationship with the objective variable, *Y*, and compared to ETR, GBR, and RFR models. A cross-correlation analysis between *Xk* (input parameters like Group-1, Group-2, and Group-3) and *Y* was conducted to examine the links between each component and coffee yield data of Sln3, Sln5B, and C×R. The Pearson correlation coefficient (*R*) was calculated using the training data to examine the associations between each parameter *Xk* and *Y*.

Table 2. Three optimal subsets of features selected for the prediction model

Models	Features Selected	Number of Features
Group-1	CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	11
Group-2	CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	8
Group-3	CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	15

Table 3. Model design parameters considered based on Group-1, Group-2, and Group-3 subsets of features

Model	Design Parameters	State of Parameters
ETR	n_estimaters	100, 200, 300, 400, 500, 600, 700, 800
	number of trees	optimal value = 600
	Min_sample_split	Int, Default = 2
	Min_sample_leaf	Int, Default = 1
	Max_features	Int, Group-1, 2, and 3 subsets, Default = 1.0
	Random_state	Int, Default = none
GBR	n_estimaters	100, 200, 300, 400, 500, 600, 700, 800
	(number of trees)	(optimal value = 600)
	Loss	Default: Squared Error.
	Learning Rate	Float, Default = 0.1
	Sub Sample	Float, Default = 1.0
	Criterion	Default = 'Friedman_mse'
	Min_sample_split	Int, Default = 2
	Min_sample_ leaf	Int, Default = 1
	Max_depth	Int/None, Default = 3
	Random_State	Int, Default = none
	Max_Features	Int, Group-1, 2, and 3 subsets, Default = None
RFR	n_estimaters	50, 100, 150, 200, 250, 300, 350, 400
	number of trees	optimal value = 200
	Max_leaf_node	Int, Default = 5
	Criterion	Default = 'Squared Error'
	Max_depth	Int, Default = none
	Min_sample_Split	Int, Default = 2
	Min_sample_leaf	Int, Default = 1
	Max_features	Int, Default = 1, Group-1, 2, and 3 subsets
	Max_leaf_node	Int, Default = none
	Bootstrap	Bool, Default = True
	Surrogate	On, Sample with replacement
DTR	Random_state	Int, Default = none
	Splitter	Best
	Max_depth	Int, Default = none
	Min_sample_split	Int, Default = 2
	Min_sample_leaf	Int, Default = 1
	Max_features	Int, Group-1, 2, and 3 subsets
	Max_leaf nodes	Int, Default = none

For model construction, evaluation/selection, and testing, measured data were individually divided into training (60%, n = 115 samples), validation (10%, n = 20 samples), and testing (30%, n = 57 samples). C×R, Sln3, and Sln5B datasets' descriptive statistics for the agro-ecological model's input parameters for training, validation, and

testing are provided in Table 2, Table 3, and Eqs. (3) – (6). Multivariate feature selection, which uses univariate statistical tests to choose the best yield-affecting attributes, was used. SelectKBest was used to pre-process an estimator by eliminating all except the *K* highest-scoring features for multivariate prediction.

In total, 32,768 distinct combinations of input features were created and analyzed based on univariate and multivariate feature selection methods to select the best prominent features, leading to better yield. Based on analysis and a feature importance function considered based on the correlation matrix, three optimal groups of features were grouped into Group-1, Group-2, and Group-3 input parameters and given as inputs to develop the prediction model to analyze the yield output. DTR models were employed in this study, with varying R^2 values for each group of feature combinations. Since each agro-ecological component was expected to affect coffee productivity, three optimum DTR models were created, which are labeled as Group-1, Group-2, and Group-3, taking into account the *K* feature values as subsets of 1, 2, 3, 4... 15. Table 2 gives the detailed information about selected input subsets for the prediction.

When constructing the choice tree ensemble, the ETR model used an approach known as *n_estimaters*, which set the number of trees to 1,000. To make an impartial comparison to the DTR model, exploratory and response correlations were regressed between the data on the components of the agro-ecology studied during the testing phase and the data on coffee production. By repeatedly experimenting with ensembles with numbers ranging from 100 to 600 in increments of one-fold, *n_estimator* was maximized, which is an important parameter. In this instance, it was discovered that 500 trees with *leaf* = 1 and *fboot* = 1 formed the best ETR model. RFR and GBR models were created for the same set of predictors (Group-1, Group-2, and Group-3) for further comparison. Interestingly, the DTR model from Group-3 performed better than the DTR models from Group-1 and Group-2. Additionally, compared to other models built utilizing Group-1, Group-2, and Group-3, the best DTR model from Group-3 performed well. The number of training/validation/testing data points in each predictor matrix was *n*. The objective criteria allowed for monitoring R^2 and RMSE in each trial and evaluating models on testing datasets. Table 3 lists the current study's four model design parameters.

2.5 Model Performance Evaluation

In the test phase, the projected and measured yields were compared using statistical performance measures to determine the accuracy of the DTR, ETR, GBR, and RFR models when applied to the coffee yield prediction issue. R^2, MAE, MSE, and RMSE were analyzed as following:

$$R^2 = 1 - \frac{SSr}{SSm} \tag{3}$$

where, SSr means Squared regression line sum error and SSm means Squared mean line sum error.

$$MAE = \frac{1}{N}\sum_{i=1}^{i=N}(|y_i - \hat{y}_i|) \tag{4}$$

$$MSE = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(y_i - \hat{y}_i)^2} \tag{6}$$

where, there are *N* anticipated values. The *i*-th data's real true value is represented by y_i. And $\hat{y}_i$ is the *i*-th data's anticipated value.

Since one metric only stresses a single feature of error characteristics, many performance metrics are generally needed to evaluate model performance (Chai & Draxler, 2014). Though widely used as goodness-of-fit statistics, the correlation coefficient and RMSE are considered suboptimal measures of model accuracy due to their oversensitivity to outliers and insensitivity to additive and proportional differences between model prediction and observations (Legates & McCabe, 2013; Legates & McCabe Jr, 1999; Willmott et al., 2012). Hence, in the present work, model performances were evaluated based on R^2 and RMSE, which have seen extensive use as goodness-of-fit metrics. Their insensitivity to additive and proportional variations between model predictions and observations and their sensitivity to extreme values (outliers) make them less than ideal measurements of model correctness. Table 4, Table 5, and Table 6 show the agro-ecological input parameters and descriptive statistics for training, validation, and testing periods of coffee crop yield (*Y*) for the C×R, Sln3, and Sln5B coffee types.

Table 4. Agro-ecological input parameters and descriptive statistics for training, validation, and testing periods of coffee crop yield for the C×R coffee type

Statistics	C×R	CLR Incidence	Tmax (°C)	Tmin (°C)	Rh (%)	Rf (cm)	OC (kg ha⁻¹)	P (kg ha⁻¹)	K (kg ha⁻¹)
Training	Mean	17.31	28.45	18.55	83.36	1761.36	6.70	48.21	371.30
	Standard deviation	21.36	3.32	2.01	7.66	1221.08	15.57	33.87	156.62
	Median	8.31	28.80	19.12	85.33	1865.20	4.00	40.00	380.00
	Minimum	0.00	22.90	12.56	56.93	27.60	2.70	1.00	63.00
	Maximum	75.30	34.67	23.78	94.10	3884.20	155.00	121.00	700.00
	Skewness	1.34	0.08	-0.59	-1.35	0.26	8.06	0.32	0.04
	Kurtosis	0.55	-1.05	0.55	1.78	-0.86	70.94	-1.38	-0.80
Validation	Mean	24.59	28.04	18.59	83.90	2009.02	4.45	46.22	387.65
	Standard deviation	24.65	2.89	2.08	7.50	1243.32	1.20	33.78	216.92
	Median	15.21	28.96	19.06	84.56	2046.15	4.14	38.50	393.00
	Minimum	0.00	23.68	12.56	57.79	225.40	2.70	2.00	102.00
	Maximum	79.69	33.34	23.14	93.37	3859.80	6.60	95.00	950.00
	Skewness	0.94	-0.01	-0.88	-2.07	0.11	0.35	0.07	0.76
	Kurtosis	-0.41	-0.96	2.44	5.46	-1.05	-0.91	-1.65	0.39
Testing	Mean	18.39	29.12	17.88	79.87	1810.36	4.40	45.60	412.84
	Standard deviation	19.19	2.83	2.02	9.47	978.52	1.08	28.56	193.07
	Median	13.35	29.12	18.00	80.60	1871.00	4.00	41.00	400.00
	Minimum	0.00	23.21	13.97	56.93	27.60	2.70	2.00	75.00
	Maximum	71.51	33.83	23.78	94.10	3884.20	5.82	114.00	915.00
	Skewness	1.21	-0.09	0.02	-0.92	0.17	0.12	0.47	0.43
	Kurtosis	0.61	-0.49	-0.18	0.37	0.18	-1.54	-0.81	-0.33
Statistics	**C×R**	**pH**	**Sp (m)**	**Ag (yr old)**	**SSmin (h)**	**SSmax (h)**	**Vp (%)**	**Dp (°C)**	**Y (kg ha⁻¹)**
Training	Mean	5.77	70.40	30.36	3.16	7.77	19.69	16.65	313.59
	Standard deviation	0.52	99.78	18.05	0.74	0.71	1.77	1.22	55.22
	Median	5.80	25.00	30.00	3.19	8.12	19.90	17.16	330.00
	Minimum	4.00	25.00	4.00	2.12	6.31	17.13	15.00	210.00
	Maximum	7.00	650.00	85.00	4.48	8.89	21.98	18.29	400.00
	Skewness	-0.36	3.01	0.44	0.29	-0.48	-0.26	-0.18	-0.22
	Kurtosis	0.81	10.78	-0.15	-1.17	-1.20	-1.54	-1.60	-1.04
Validation	Mean	5.56	45.65	32.55	3.15	7.67	19.41	16.42	313.
	Standard deviation	0.58	52.21	17.59	0.81	0.70	1.90	1.18	50.39
	Median	5.60	25.00	32.00	3.14	8.04	19.81	17.03	334.00
	Minimum	4.50	25.00	8.00	2.12	6.66	17.13	15.03	218.00
	Maximum	6.50	250.00	67.00	4.47	8.59	21.83	18.24	395.00
	Skewness	-0.20	3.27	0.31	0.31	-0.25	-0.01	0.07	-0.42
	Kurtosis	-0.92	10.09	-1.09	-1.28	-1.59	-1.74	-1.59	-0.8
Testing	Mean	5.85	79.95	30.05	3.22	7.71	19.33	16.50	307.58
	Standard deviation	0.56	96.38	21.65	0.77	0.74	1.78	1.19	53.17
	Median	5.80	25.00	27.00	3.33	8.15	19.85	17.09	300.00
	Minimum	4.70	25.00	4.00	2.21	6.17	17.13	15.04	218.00
	Maximum	7.50	350.00	85.00	4.47	8.85	21.90	18.28	398.00
	Skewness	0.19	1.61	1.12	0.17	-0.46	0.01	0.01	-0.01

Table 5. Agro-ecological input parameters and descriptive statistics for training, validation, and testing periods of coffee crop yield for the Sln3 coffee type

Statistics	Sln3	CLR Incidence	Tmax (°C)	Tmin (°C)	Rh (%)	Rf (cm)	OC (kg ha⁻¹)	P (kg ha⁻¹)	K (kg ha⁻¹)
Training	Mean	5.45	28.45	18.55	83.36	1761.36	6.70	48.21	371.30
	Standard deviation	7.74	3.32	2.01	7.66	1221.08	15.57	33.87	156.62
	Median	2.56	28.80	19.12	85.33	1865.20	4.00	40.00	380.00
	Minimum	0.00	22.90	12.56	56.93	27.60	2.70	1.00	63.00
	Maximum	35.91	34.67	23.78	94.10	3884.20	155.00	121.00	700.00

	Skewness	2.23	0.08	-0.59	-1.35	0.26	8.06	0.32	0.04
	Kurtosis	4.82	-1.05	0.55	1.78	-0.86	70.94	-1.38	-0.80
Validation	Mean	6.34	28.04	18.59	83.90	2009.02	4.45	46.22	387.65
	Standard deviation	6.90	2.89	2.08	7.50	1243.32	1.20	33.78	216.92
	Median	4.45	28.96	19.06	84.56	2046.15	4.14	38.50	393.00
	Minimum	0.00	23.68	12.56	57.79	225.40	2.70	2.00	102.00
	Maximum	27.87	33.34	23.14	93.37	3859.80	6.60	95.00	950.00
	Skewness	1.63	-0.01	-0.88	-2.07	0.11	0.35	0.07	0.76
	Kurtosis	2.73	-0.96	2.44	5.46	-1.05	-0.91	-1.65	0.39
Testing	Mean	9.04	29.12	17.88	79.87	1810.36	4.40	45.60	412.84
	Standard deviation	11.23	2.83	2.02	9.47	978.52	1.08	28.56	193.07
	Median	2.80	29.12	18.00	80.60	1871.00	4.00	41.00	400.00
	Minimum	0.00	23.21	13.97	56.93	27.60	2.70	2.00	75.00
	Maximum	37.27	33.83	23.78	94.10	3884.20	5.82	114.00	915.00
	Skewness	1.29	-0.09	0.02	-0.92	0.17	0.12	0.47	0.43
	Kurtosis	0.49	-0.49	-0.18	0.37	0.18	-1.54	-0.81	-0.33
Statistics Sln3		**pH**	**Sp (m)**	**Ag (yr old)**	**SSmin (h)**	**SSmax (h)**	**Vp (%)**	**Dp (°C)**	**Y (kg ha⁻¹)**
Training	Mean	5.77	70.40	30.36	3.16	7.77	19.69	16.65	359.20
	Standard deviation	0.52	99.78	18.05	0.74	0.71	1.77	1.22	39.99
	Median	5.80	25.00	30.00	3.19	8.12	19.90	17.16	350.00
	Minimum	4.00	25.00	4.00	2.12	6.31	17.13	15.00	294.00
	Maximum	7.00	650.00	85.00	4.48	8.89	21.98	18.29	400.00
	Skewness	-0.36	3.01	0.44	0.29	-0.48	-0.26	-0.18	-0.46
	Kurtosis	0.81	10.78	-0.15	-1.17	-1.20	-1.54	-1.60	-1.32
Validation	Mean	5.56	45.65	32.55	3.15	7.67	19.41	16.42	367.75
	Standard deviation	0.58	52.21	17.59	0.81	0.70	1.90	1.18	34.40
	Median	5.60	25.00	32.00	3.14	8.04	19.81	17.03	389.00
	Minimum	4.50	25.00	8.00	2.12	6.66	17.13	15.03	295.00
	Maximum	6.50	250.00	67.00	4.47	8.59	21.83	18.24	400.00
	Skewness	-0.20	3.27	0.31	0.31	-0.25	-0.01	0.07	-0.72
	Kurtosis	-0.92	10.09	-1.09	-1.28	-1.59	-1.74	-1.59	-0.67
Testing	Mean	5.85	79.95	30.05	3.22	7.71	19.33	16.50	355.21
	Standard deviation	0.56	96.38	21.65	0.77	0.74	1.78	1.19	44.49
	Median	5.80	25.00	27.00	3.33	8.15	19.85	17.09	388.00
	Minimum	4.70	25.00	4.00	2.21	6.17	17.13	15.04	294.00
	Maximum	7.50	350.00	85.00	4.47	8.85	21.90	18.28	400.00
	Skewness	0.19	1.61	1.12	0.17	-0.46	0.01	0.01	-0.35
	Kurtosis	0.06	0.95	0.93	-1.35	-1.38	-1.61	-1.61	-1.66

Table 6. Agro-ecological input parameters and descriptive statistics for training, validation, and testing periods of coffee crop yield for the Sln5B coffee type

Statistics Sln5B		**CLR Incidence**	**Tmax (°C)**	**Tmin (°C)**	**Rh (%)**	**Rf (cm)**	**OC (kg ha⁻¹)**	**P (kg ha⁻¹)**	**K (kg ha⁻¹)**
Training	Mean	3.86	28.45	18.55	83.36	1761.36	6.70	48.21	371.30
	Standard deviation	3.83	3.32	2.01	7.66	1221.08	15.57	33.87	156.62
	Median	2.40	28.80	19.12	85.33	1865.20	4.00	40.00	380.00
	Minimum	0.00	22.90	12.56	56.93	27.60	2.70	1.00	63.00
	Maximum	14.70	34.67	23.78	94.10	3884.20	155.00	121.00	700.00
	Skewness	1.18	0.08	-0.59	-1.35	0.26	8.06	0.32	0.04
	Kurtosis	0.61	-1.05	0.55	1.78	-0.86	70.94	-1.38	-0.80
Validation	Mean	5.28	28.04	18.59	83.90	2009.02	4.45	46.22	387.65
	Standard deviation	3.81	2.89	2.08	7.50	1243.32	1.20	33.78	216.92
	Median	4.54	28.96	19.06	84.56	2046.15	4.14	38.50	393.00
	Minimum	0.58	23.68	12.56	57.79	225.40	2.70	2.00	102.00
	Maximum	12.48	33.34	23.14	93.37	3859.80	6.60	95.00	950.00
	Skewness	0.48	-0.01	-0.88	-2.07	0.11	0.35	0.07	0.76
	Kurtosis	-1.03	-0.96	2.44	5.46	-1.05	-0.91	-1.65	0.39

Testing	Mean	4.41	29.12	17.88	79.87	1810.36	4.40	45.60	412.84
	Standard deviation	3.42	2.83	2.02	9.47	978.52	1.08	28.56	193.07
	Median	3.85	29.12	18.00	80.60	1871.00	4.00	41.00	400.00
	Minimum	0.00	23.21	13.97	56.93	27.60	2.70	2.00	75.00
	Maximum	12.48	33.83	23.78	94.10	3884.20	5.82	114.00	915.00
	Skewness	0.61	-0.09	0.02	-0.92	0.17	0.12	0.47	0.43
	Kurtosis	-0.56	-0.49	-0.18	0.37	0.18	-1.54	-0.81	-0.33
Statistics Sln5B		**pH**	**Sp (m)**	**Ag (yr old)**	**SSmin (h)**	**SSmax (h)**	**Vp (%)**	**Dp (℃)**	**Y (kg ha⁻¹)**
Training	Mean	5.77	70.40	30.36	3.16	7.77	19.69	16.65	339.10
	Standard deviation	0.52	99.78	18.05	0.74	0.71	1.77	1.22	29.19
	Median	5.80	25.00	30.00	3.19	8.12	19.90	17.16	330.00
	Minimum	4.00	25.00	4.00	2.12	6.31	17.13	15.00	264.00
	Maximum	7.00	650.00	85.00	4.48	8.89	21.98	18.29	400.00
	Skewness	-0.36	3.01	0.44	0.29	-0.48	-0.26	-0.18	-0.41
	Kurtosis	0.81	10.78	-0.15	-1.17	-1.20	-1.54	-1.60	-1.02
Validation	Mean	5.56	45.65	32.55	3.15	7.67	19.41	16.42	347.75
	Standard deviation	0.58	52.21	17.59	0.81	0.70	1.90	1.18	31.40
	Median	5.60	25.00	32.00	3.14	8.04	19.81	17.03	369.00
	Minimum	4.50	25.00	8.00	2.12	6.66	17.13	15.03	298.00
	Maximum	6.50	250.00	67.00	4.47	8.59	21.83	18.24	400.00
	Skewness	-0.20	3.27	0.31	0.31	-0.25	-0.01	0.07	-0.32
	Kurtosis	-0.92	10.09	-1.09	-1.28	-1.59	-1.74	-1.59	-0.47
Testing	Mean	5.85	79.95	30.05	3.22	7.71	19.33	16.50	365.21
	Standard deviation	0.56	96.38	21.65	0.77	0.74	1.78	1.19	44.49
	Median	5.80	25.00	27.00	3.33	8.15	19.85	17.09	378.00
	Minimum	4.70	25.00	4.00	2.21	6.17	17.13	15.04	284.00
	Maximum	7.50	350.00	85.00	4.47	8.85	21.90	18.28	400.00
	Skewness	0.19	1.61	1.12	0.17	-0.46	0.01	0.01	-0.25
	Kurtosis	0.06	0.95	0.93	-1.35	-1.38	-1.61	-1.61	-1.56

3. Results

Scatterplots were used to visually examine the degree of agreement between actual and predicted yield data during the testing phase to reduce all features to a single scale without modifying the range of values in the stochastic models. In this study, the features were chosen using a multivariate feature selection strategy considering the feature significance function and a correlation matrix. The conventional scalar fine-tuning method was also used. Then three ideal models were selected and classified into three sets of parameters while assessing and creating the model into Group-1, Group-2, and Group-3 input parameters to produce the prediction model to examine yield output. Scatterplots show R^2, RMSE, actual and predicted yield, and prediction error.

3.1 ETR

ET, sometimes known as extremely randomized trees, is a regression model similar to RF. The features are randomly divided. In the present work, 100 DTs from various important categories were used to develop the ETR model, which combined the training and test datasets. The ETR model was used to forecast coffee yield with variable agro-ecological factors for three different groups, as well as the outcomes of the testing phase assessment for different split ratios. Table 7, Table 8, and Table 9 show the ratios of several performance indicators, including R^2, MAE, MSE, and RMSE. Boldface indicates the optimal values obtained for the C×R, Sln3, and Sln5B datasets.

By comparing the actual yield with the measured testing-phase yield graphically, it was found that Group-1 for C×R had the highest R^2 (0.98), Group-3 for Sln3 had an R^2 of 0.98, and Group-3 of Sln5B datasets for 70:30 split ratios had an R^2 of 0.97. The ETR model performance for the C×R dataset, as indicated in Table 7, demonstrated high accuracy for Group-1 input parameters, with an R^2 of 0.98 and an RMSE of 8.61 kg ha^{-1}. For the Sln3 dataset presented in Table 8, the model was accurate for Group-3 parameters, with an R^2 of 0.98 and an RMSE of 8.27 kg ha^{-1}. Similarly, the Sln5B dataset in Table 9 showed accuracy for the same Group-3 parameters, with an R^2 of 0.97 and an RMSE of 7.79 kg ha^{-1}. Figure 2 shows the measured and actual yield visual representation through scatterplots of C×R, Sln3, and Sln5B for Groups 1, 2, and 3, which served as input parameters for the ETR-based stochastic models. Scatterplots depict the highest expected and actual coffee yields during experimentation (70:30 splits).

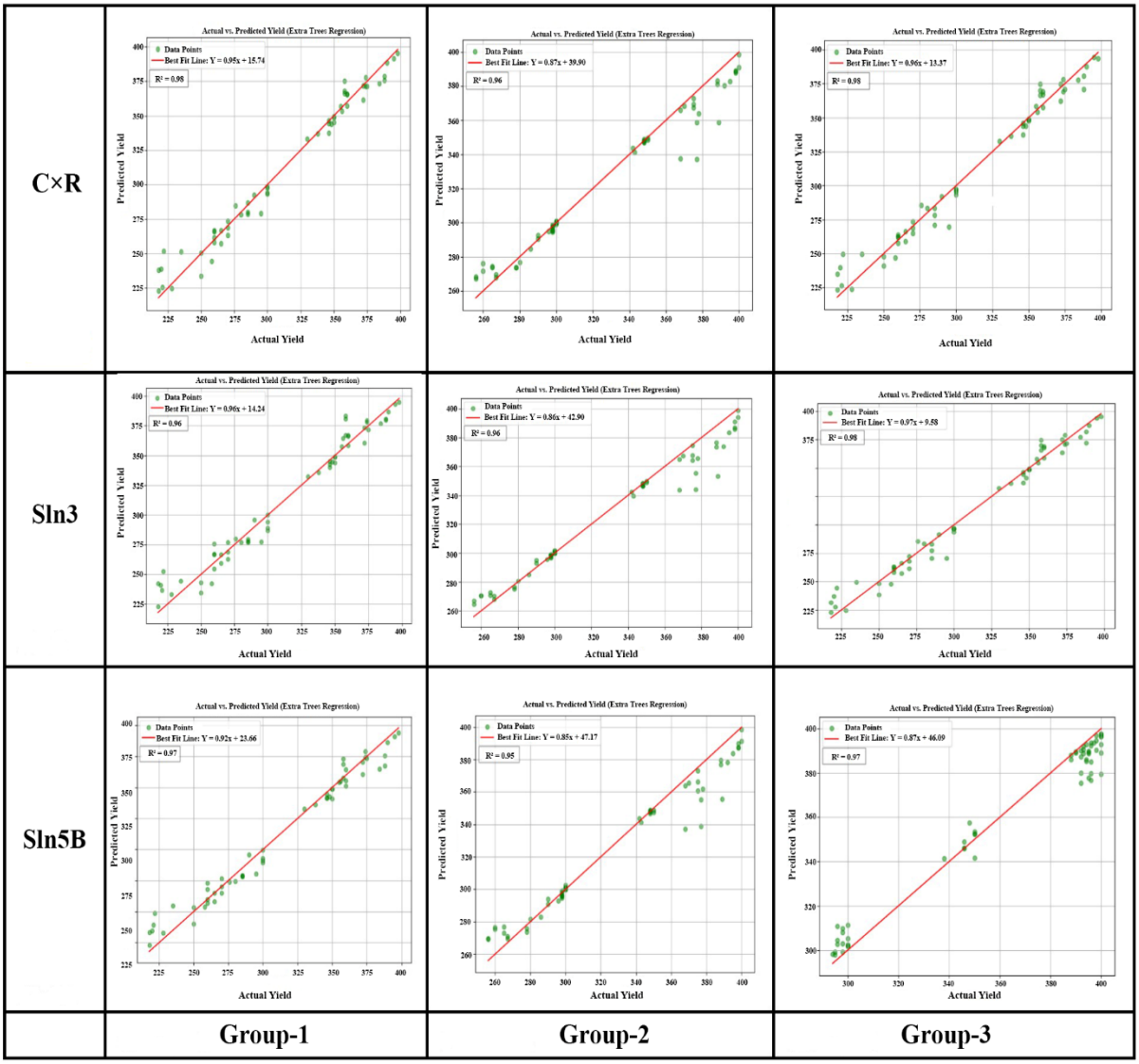

Figure 2. Measured and actual yield visual representation of C×R, Sln3, and Sln5B for Groups 1, 2, and 3 (ETR), reflecting the greatest predicted yield displayed along the Y-axis and the actual coffee yields displayed along the X-axis during the testing period for 70:30 splits

Table 7. Performance of the ETR model for coffee yield prediction (C×R)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.98	5.63	68.41	8.27
	80:20	0.98	5.63	68.41	8.27
	70:30	**0.98**	**6.27**	**74.17**	**8.61**
	60:40	0.97	6.12	74.98	8.66
	50:50	0.96	7.72	111.35	10.55
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.96	6.13	86.12	9.28
	80:20	0.95	6.85	123.79	11.13
	70:30	0.96	5.86	98.54	9.93
	60:40	0.95	5.51	85.41	9.24
	50:50	0.94	6.78	114.91	10.72
Group-3 – CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.97	6.49	85.16	9.23
	80:20	0.98	6.24	80.02	8.95
	70:30	0.97	6.39	76.41	8.74
	60:40	0.97	6.98	92.35	9.61
	50:50	0.95	8.04	130.22	11.41

Table 8. Performance of the ETR model for coffee yield prediction (Sln3)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.97	8.17	111.98	10.58
	80:20	0.96	8.83	138.43	11.77
	70:30	0.96	7.83	105.96	10.29
	60:40	0.96	8.19	114.86	10.72
	50:50	0.94	9.46	154.81	12.44
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.98	4.11	41.96	6.48
	80:20	0.96	6.23	98.60	9.93
	70:30	0.96	5.80	94.96	9.74
	60:40	0.96	4.99	75.17	8.67
	50:50	0.95	5.71	96.39	9.82
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.98	4.41	32.62	5.71
	80:20	0.98	4.41	32.62	5.71
	70:30	**0.98**	**6.16**	**68.44**	**8.27**
	60:40	0.98	4.94	41.91	6.47
	50:50	0.90	9.93	179.66	13.40

Table 9. Performance of the ETR model for coffee yield prediction (Sln5B)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.97	7.93	103.19	10.16
	80:20	0.97	7.80	99.27	9.96
	70:30	0.97	6.86	82.83	9.10
	60:40	0.97	6.75	77.61	8.81
	50:50	0.94	9.24	163.29	12.78
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.97	6.39	82.31	9.07
	80:20	0.95	7.12	120.95	11.00
	70:30	0.95	6.72	114.80	10.71
	60:40	0.95	5.87	94.47	9.72
	50:50	0.94	6.79	113.96	10.68
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.98	4.57	35.74	5.98
	80:20	0.94	7.65	90.96	9.54
	70:30	**0.97**	**5.90**	**60.70**	**7.79**
	60:40	0.96	6.07	68.03	8.25
	50:50	0.89	10.92	207.63	14.41

3.2 GBR

The base estimator is considered an RF, and 100 weak learners were used to boost the current case. At each stage, the number of weak learners was raised to compensate for the weak learners present. Errors in the combined model may be discovered using gradients. The GBR model was used to forecast coffee yield with variable agro-ecological factors for three different groups and the outcomes of the testing phase assessment for different split ratios. Table 10, Table 11 and Table 12 show the ratios of several performance indicators, including R^2, MAE, MSE, and RMSE. Boldface indicates the optimal values obtained for the C×R, Sln3, and Sln5B datasets.

Table 10. Performance of the GBR model for coffee yield prediction (C×R)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.95	9.53	156.94	12.53
	80:20	0.94	10.25	186.77	13.67
	70:30	0.95	9.15	152.76	12.36
	60:40	0.95	8.14	137.19	11.71
	50:50	0.90	12.01	272.27	16.50
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.98	4.82	42.56	6.52
	80:20	0.94	5.95	132.95	11.53
	70:30	0.94	5.95	132.95	11.53
	60:40	0.95	5.33	88.49	9.41
	50:50	0.94	6.12	105.18	10.26
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.93	7.81	106.93	10.34
	80:20	0.93	7.81	106.93	10.34
	70:30	**0.96**	**8.32**	**126.05**	**11.23**
	60:40	0.92	8.80	133.51	11.55
	50:50	0.82	12.17	329.67	18.16

Table 11. Performance of the GBR model for coffee yield prediction (Sln3)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.95	9.53	156.94	12.53
	80:20	0.94	10.25	186.77	13.67
	70:30	0.95	9.15	152.76	12.36
	60:40	0.95	8.14	137.19	11.71
	50:50	0.90	12.01	272.27	16.50
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.98	4.82	42.56	6.52
	80:20	0.94	5.95	132.95	11.53
	70:30	0.94	5.95	132.95	11.53
	60:40	0.95	5.33	88.49	9.41
	50:50	0.94	6.12	105.18	10.26
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.93	7.81	106.93	10.34
	80:20	0.93	7.81	106.93	10.34
	70:30	**0.96**	**8.32**	**126.05**	**11.23**
	60:40	0.92	8.80	133.51	11.55
	50:50	0.82	12.17	329.67	18.16

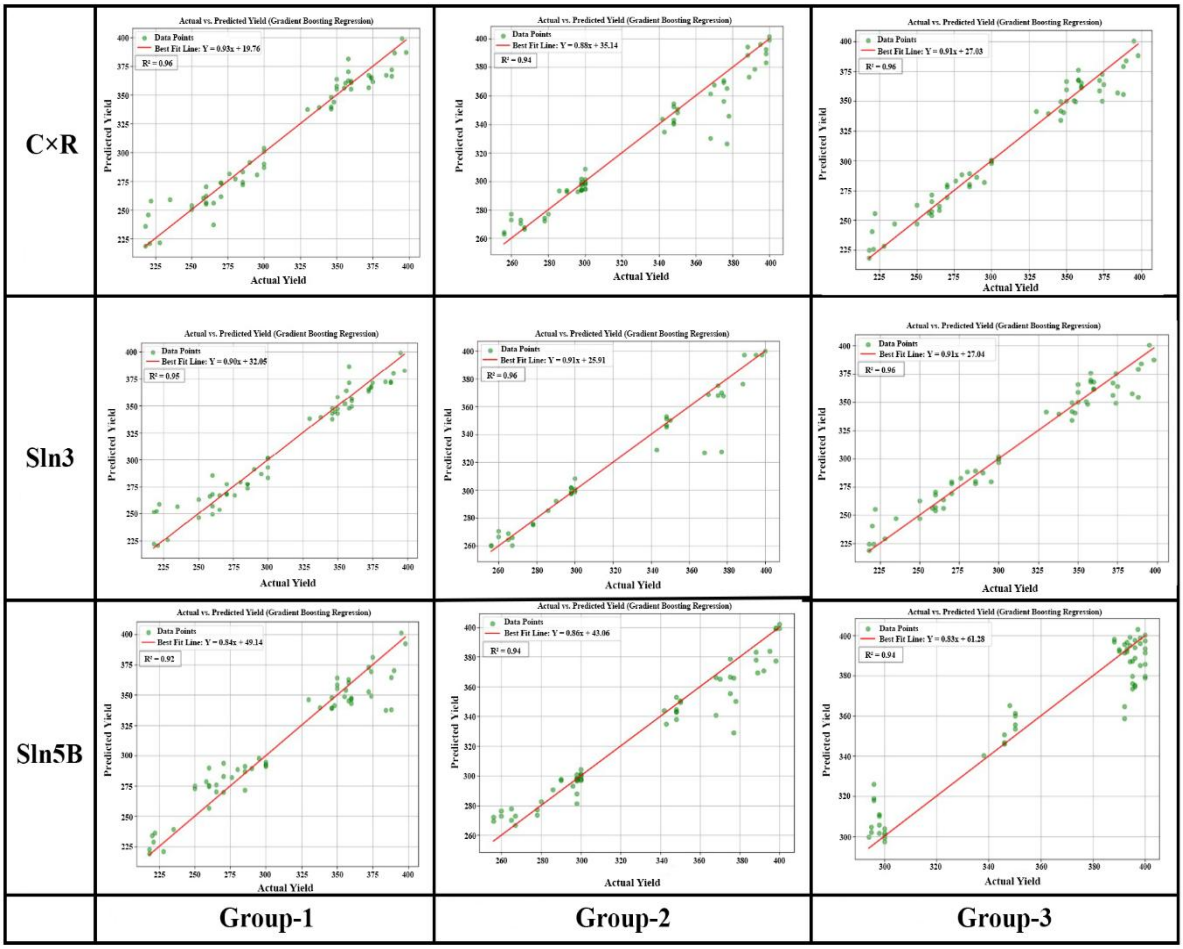

Figure 3. Measured and actual yield visual representation of C×R, Sln3, and Sln5B for Groups 1, 2, and 3 (GBR), reflecting the greatest predicted yield displayed along the Y-axis and the actual coffee yields displayed along the X-axis during the testing period for 70:30 splits

The maximum R^2 (0.96) was found for Group-3 for C×R, an R^2 of 0.96 for Sln3, and an R^2 of 0.94 for Group-2 of Sln5B datasets for 70:30 split ratios, according to a graphic comparison of the actual and measured testing phase yields (Table 10, Table 11, and Table 12). As shown in Table 10, the GBR model's performance on the C×R dataset showed great accuracy for Group-3 input parameters, with an RMSE of 10.99 kg ha^{-1} and an R^2 of 0.96. The model was accurate for Group-3 characteristics, including CLR, an R^2 of 0.96, and an RMSE of 11.23 kg ha^{-1} for the Sln3 dataset shown in Table 11. For the identical Group-2 characteristics, the Sln5B dataset in Table

12 demonstrated accuracy with an RMSE of 11.67 kg ha^{-1} and an R^2 of 0.94. As input parameters for the GBR-based stochastic models, the C×R, Sln3, and Sln5B scatterplots for Groups 1, 2, and 3 visually depict the measured and actual yields in Figure 3. The greatest anticipated and actual coffee yields over the trial period are shown in scatterplots (70:30 splits).

Table 12. Performance of the GBR model for coffee yield prediction (Sln5B)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.98	6.08	58.31	7.64
	80:20	0.94	10.04	198.91	14.10
	70:30	0.92	11.27	228.12	15.10
	60:40	0.94	8.65	164.31	12.82
	50:50	0.90	10.73	266.72	16.33
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.96	6.64	85.39	9.24
	80:20	0.91	8.45	212.69	14.58
	70:30	**0.94**	**7.65**	**136.16**	**11.67**
	60:40	0.92	7.33	155.39	12.47
	50:50	0.92	7.52	143.46	11.98
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.91	8.65	134.43	11.59
	80:20	0.91	9.37	145.34	12.06
	70:30	0.92	8.91	149.38	12.22
	60:40	0.94	7.20	98.43	9.92
	50:50	0.85	11.62	286.24	16.92

3.3 RFR

In this study, *n_estimators* = 400, the number of DTs in the forest was constructed, and the hyper parameter was also increased to improve the performance of the model. It trained several DTs on a random collection of data and attributes. The final forecast was the average of all predictions. The RFR model was used to forecast coffee yield with variable agro-ecological factors for three different groups and the outcomes of the testing phase assessment for different split ratios. Table 13, Table 14, and Table 15 show the ratios of several performance indicators, including R^2, MAE, MSE, and RMSE. Boldface indicates the optimal values obtained for the C×R, Sln3, and Sln5B datasets.

Table 13. Performance of the RFR model for coffee yield prediction (C×R)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.91	11.87	295.96	17.20
	80:20	0.91	11.87	295.96	17.20
	70:30	0.91	11.23	272.10	16.50
	60:40	0.89	12.11	296.48	17.22
	50:50	0.83	13.32	444.87	21.09
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.85	14.25	353.84	18.81
	80:20	0.91	9.94	206.72	14.38
	70:30	**0.91**	**9.18**	**199.45**	**14.12**
	60:40	0.91	8.04	160.17	12.66
	50:50	0.89	9.96	197.78	14.06
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.84	16.12	558.66	23.64
	80:20	0.88	12.05	389.02	19.72
	70:30	0.90	10.40	297.69	17.25
	60:40	0.89	11.00	290.46	17.04
	50:50	0.74	14.73	709.42	26.63

Table 14. Performance of the RFR model for coffee yield prediction (Sln3)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.91	13.25	294.53	17.16
	80:20	0.91	12.00	295.89	17.20
	70:30	0.91	11.56	274.57	16.57
	60:40	0.88	12.07	317.43	17.82
	50:50	0.81	14.99	504.99	22.47
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.94	8.96	144.96	12.04
	80:20	0.94	8.44	146.33	12.10
	70:30	**0.95**	**7.31**	**114.67**	**10.71**

	60:40	0.94	7.01	116.87	10.81
	50:50	0.92	8.41	149.67	12.23
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.96	5.64	55.68	7.46
	80:20	0.96	5.64	55.68	7.46
	70:30	0.89	11.92	345.88	18.60
	60:40	0.90	10.10	167.98	12.96
	50:50	0.78	15.56	409.06	20.23

Table 15. Performance of the RFR model for coffee yield prediction (Sln5B)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag.	90:10	0.94	11.06	201.43	14.19
	80:20	0.93	11.73	227.60	15.09
	70:30	0.93	11.24	205.05	14.32
	60:40	0.93	11.09	198.22	14.08
	50:50	0.86	13.24	378.49	19.45
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH.	90:10	0.90	11.62	236.70	15.39
	80:20	0.90	10.18	224.71	14.99
	70:30	0.93	8.19	148.98	12.21
	60:40	0.90	8.57	178.45	13.36
	50:50	0.89	9.92	190.01	13.78
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp.	90:10	0.96	6.13	68.69	8.29
	80:20	0.95	6.71	82.94	9.11
	70:30	**0.93**	**7.81**	**121.14**	**11.01**
	60:40	0.92	9.25	149.25	12.22
	50:50	0.74	16.73	487.74	22.08

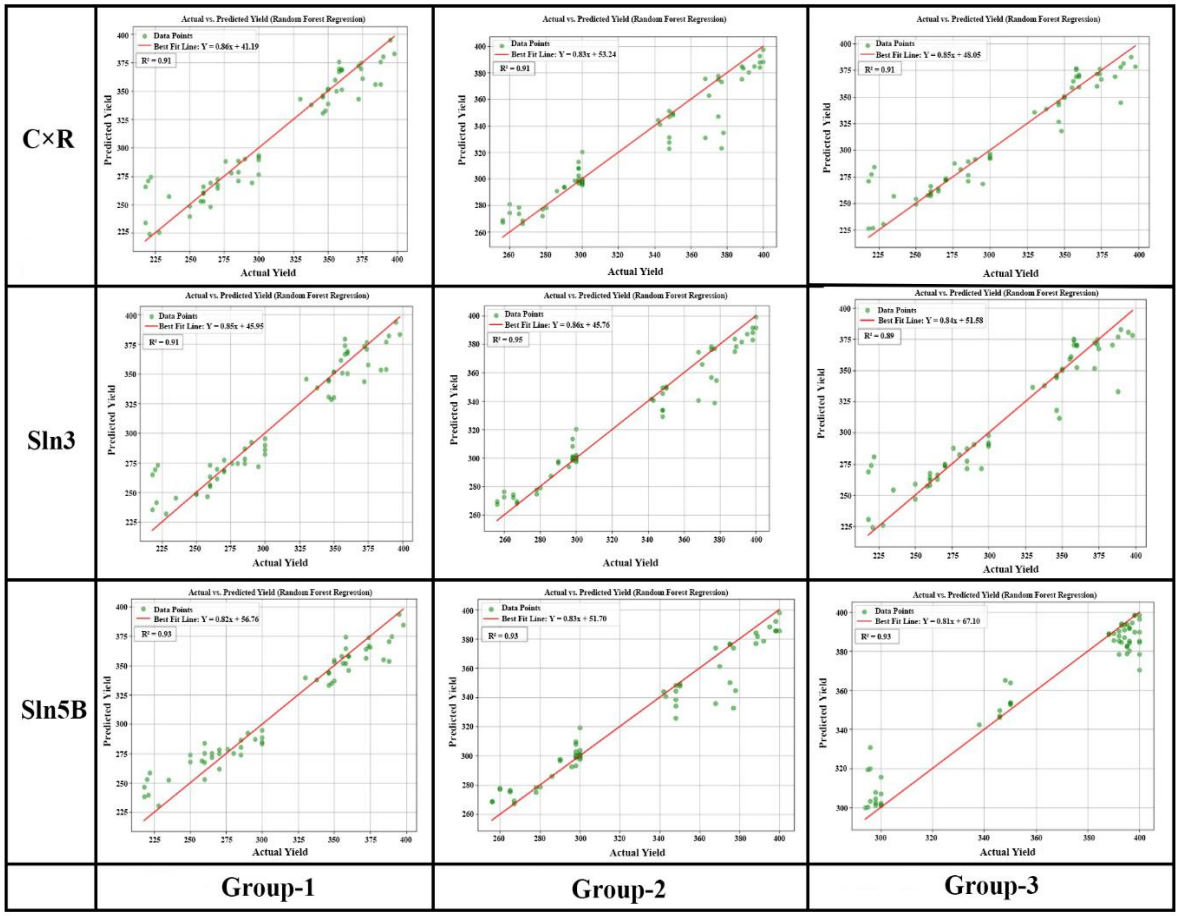

Figure 4. Measured and actual yield visual representation of C×R, Sln3, and Sln5B for Groups 1, 2, and 3 (RFR), reflecting the greatest predicted yield displayed along the Y-axis and the actual coffee yields displayed along the X-axis during the testing period for 70:30 splits

According to a graphic comparison of the actual yield and the measured testing-phase yield, the maximum R^2 (0.91) was determined for Group-2 for C×R, an R^2 of 0.95 for Sln3 for Group-2, and an R^2 of 0.93 for Group-3 of Sln5B datasets for 70:30 split ratios. Table 13 illustrates that the RFR model demonstrated high accuracy for Group-2 input parameters on the C×R dataset, with an RMSE of 14.12 kg ha^{-1} and an R^2 of 0.91. The Sln3 dataset, as illustrated in Table 14, demonstrated that the model was accurate for Group-2 characteristics, including CLR, an R^2 of 0.95, and an RMSE of 10.71 kg ha^{-1}. The Sln5B dataset in Table 15 exhibits accuracy with an RMSE of 11.01 kg ha^{-1} and an R^2 of 0.93 for the identical Group-3 characteristics. The C×R, Sln3, and Sln5B scatterplots of Groups 1, 2, and 3 serve as input parameters for the RFR-based stochastic models in Figure 4, visually representing the actual and measured yields. Scatterplots (70:30 splits) illustrate the highest anticipated and actual coffee yields during the trial period.

3.4 DTR

DTR employs a tree structure to partition data according to characteristics, with the projected numerical values included in the leaf nodes. Nodes, branches, and leaves make up the model's representation as a tree. Data is subdivided iteratively according to the values of features. The DTR model was used to forecast coffee yield with variable agro-ecological factors for three different groups and the outcomes of the testing phase assessment for different split ratios. Table 16, Table 17, and Table 18 show the ratios of several performance indicators, including R^2, MAE, MSE, and RMSE. Boldface indicates the optimal values obtained for the C×R, Sln3, and Sln5B datasets.

Table 16. Performance of the DTR model for coffee yield prediction (C×R)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.97	5.69	82.46	9.08
	80:20	0.97	5.69	82.46	9.08
	70:30	0.87	10.90	390.59	19.76
	60:40	0.91	8.79	247.86	15.74
	50:50	0.29	19.82	1896.14	43.54
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.33	25.10	1592.30	39.90
	80:20	0.84	8.33	364.28	19.09
	70:30	0.70	11.28	657.52	25.64
	60:40	0.73	9.47	510.25	22.59
	50:50	0.56	13.88	786.21	28.04
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	0.47	21.40	1812.50	42.57
	80:20	0.94	8.67	181.59	13.48
	70:30	**0.91**	**9.91**	**259.16**	**16.10**
	60:40	0.91	10.14	251.13	15.85
	50:50	0.34	18.41	1766.41	42.03

Table 17. Performance of the DTR model for coffee yield prediction (Sln3)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	-0.04	38.90	3554.60	59.62
	80:20	0.91	11.41	295.41	17.19
	70:30	0.87	13.60	404.02	20.10
	60:40	0.85	12.97	397.23	19.93
	50:50	0.89	12.04	308.21	17.56
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.84	9.15	375.05	19.37
	80:20	0.84	7.92	355.00	18.84
	70:30	0.77	9.09	506.78	22.51
	60:40	0.79	7.83	391.03	19.77
	50:50	0.60	12.00	703.92	26.53
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	1.00	1.45	3.25	1.80
	80:20	1.00	1.45	3.25	1.80
	70:30	**0.90**	**9.67**	**249.16**	**15.10**
	60:40	0.93	4.42	119.71	10.94
	50:50	0.35	15.77	1227.54	35.04

The Group-3 DTR model for C×R had the greatest R^2 (0.91), the Group-3 Sln3 model had an R^2 of 0.90, and the Group-3 Sln5B model had an R^2 of 1.0 for 70:30 split ratios, as shown in the graphic comparison of the actual yield with the measured testing-phase yield. The DTR model showed good accuracy for Group-3 input parameters on the C×R dataset, as shown in Table 16, with an RMSE of 16.10 kg ha^{-1} and an R^2 of 0.91. Table 17 shows that the Sln3 dataset proved the model's accuracy for Group-3 attributes, such as CLR, with an R^2 of 0.90 and an

RMSE of 15.10 kg ha^{-1}. Table 18 shows that the Sln5B dataset is accurate for the same Group-3 features, with an RMSE of 1.27 kg ha^{-1} and an R^2 of 1.0. Figure 5 shows the DTR-based stochastic models that use the C×R, Sln3, and Sln5B scatterplots from Groups 1, 2, and 3 as input parameters. These scatterplots show the difference between the actual and measured yields. Scatterplots showing 70:30 splits show the greatest expected and actual coffee yields throughout the experiment.

Table 18. Performance of the DTR model for coffee yield prediction (Sln5B)

Groups	Split Ratio	R^2	MAE	MSE	RMSE
Group-1: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Sp, and Ag	90:10	0.95	7.65	180.85	13.45
	80:20	0.91	12.38	304.13	17.44
	70:30	0.87	13.48	407.03	20.18
	60:40	0.90	11.36	272.32	16.56
	50:50	0.81	15.00	499.62	22.35
Group-2: CLR, Tmin, Tmax, Rh, Rf, Ag, OC, and pH	90:10	0.88	9.25	291.55	17.07
	80:20	0.84	9.08	367.23	19.16
	70:30	0.69	11.91	681.16	26.10
	60:40	0.74	9.49	492.43	22.19
	50:50	0.76	9.00	417.94	20.44
Group-3: CLR, Tmin, Tmax, Rh, Rf, OC, P, K, pH, Ag, Sp, SSmin, SSmax, Vp, and Dp	90:10	1.00	1.10	2.20	1.48
	80:20	1.00	0.95	1.77	1.33
	70:30	**1.00**	**0.83**	**1.62**	**1.27**
	60:40	0.92	3.44	139.36	11.81
	50:50	0.80	6.19	381.69	19.54

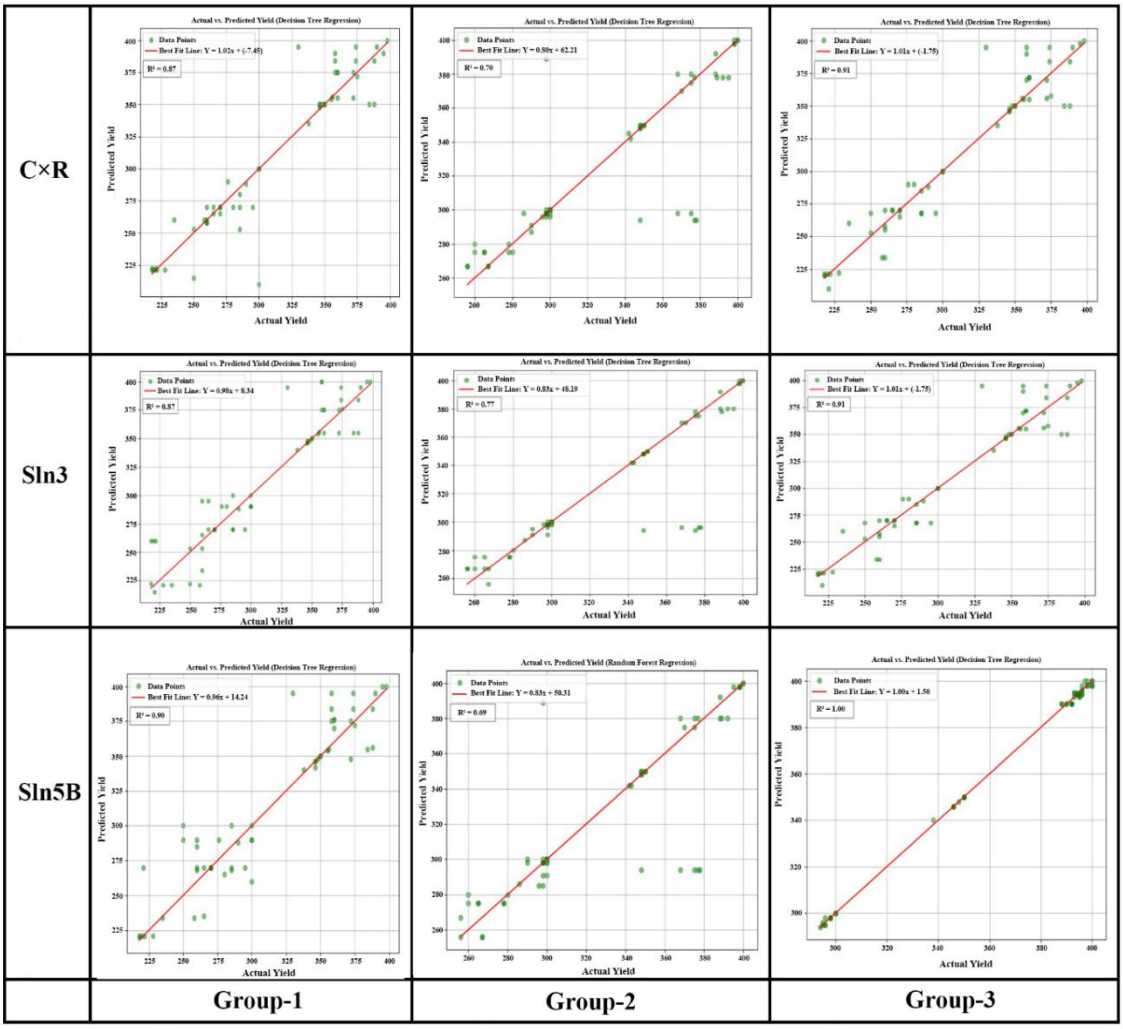

Figure 5. Measured and actual yield visual representation of C×R, Sln3, and Sln5B for Groups 1, 2, and 3 (DTR), reflecting the greatest predicted yield displayed along the Y-axis and the actual coffee yields displayed along the X-axis during the testing period for 70:30 splits

Table 19. Descriptive statistics of model errors for Group-1 input parameters during the testing period (kg ha⁻¹)

Multivariate Error Distribution	Considered Models			
	DTR	RFR	GBR	ETR
Lower quartile: p25	270.00	268.32	269.92	269.85
Median: p50	326.00	329.42	326.01	328.41
Upper quartile: p75	350.00	350.27	350.03	349.66
Maximum	400.00	396.65	400.00	398.76
Minimum	210.00	218.33	209.92	217.20
Standard deviation	53.06	50.12	53.05	52.29
Skewness	-0.20	-0.13	-0.20	-0.16
Kurtosis	-1.01	-1.20	-1.01	-1.10

Table 20. Descriptive statistics of model errors for Group-2 input parameters during the testing period (kg ha⁻¹)

Multivariate Error Distribution	Considered Models			
	DTR	RFR	GBR	ETR
Lower quartile: p25	298.00	298.38	297.92	297.93
Median: p50	300.00	326.12	300.07	333.21
Upper quartile: p75	350.00	349.67	350.02	349.60
Maximum	400.00	399.04	400.08	399.70
Minimum	256.00	262.35	255.95	263.69
Standard deviation	41.35	39.46	41.34	40.52
Skewness	0.36	0.34	0.36	0.36
Kurtosis	-1.14	-1.13	-1.14	-1.15

Table 21. Descriptive statistics of model errors for Group-3 input parameters during the testing period (kg ha⁻¹)

Multivariate Error Distribution	Considered Models			
	DTR	RFR	GBR	ETR
Lower quartile: p25	300.00	309.13	300.10	302.60
Median: p50	350.00	352.04	350.02	350.62
Upper quartile: p75	396.00	393.46	396.05	395.69
Maximum	400.00	399.85	400.15	399.81
Minimum	294.00	295.12	293.82	294.53
Standard deviation	40.79	38.42	40.78	40.09
Skewness	-0.36	-0.38	-0.36	-0.36
Kurtosis	-1.44	-1.41	-1.44	-1.44

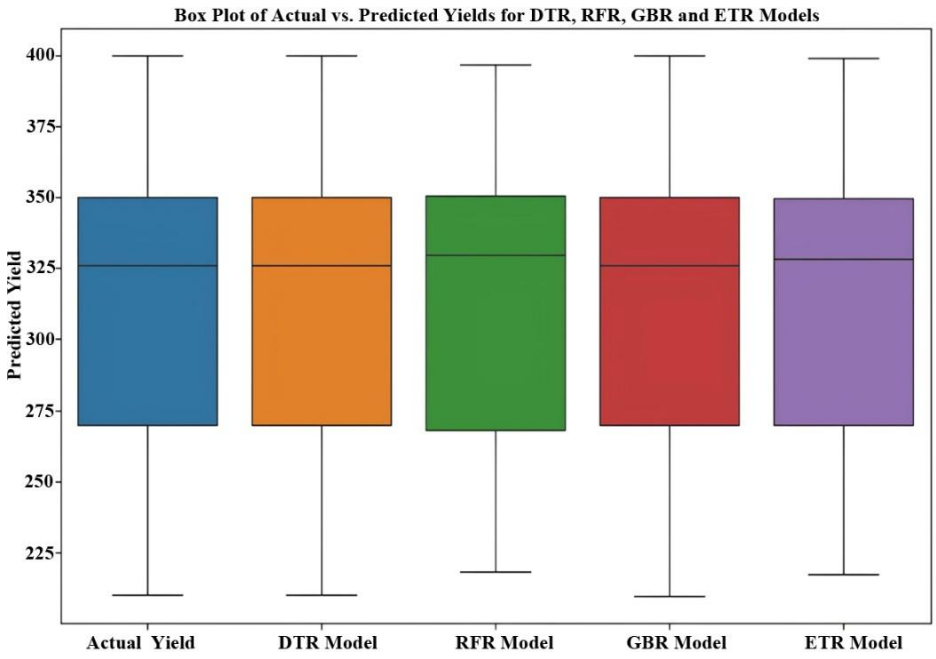

Figure 6. Anticipated coffee yield using the best four models for the best Group-1 characteristics and measured coffee yield during testing

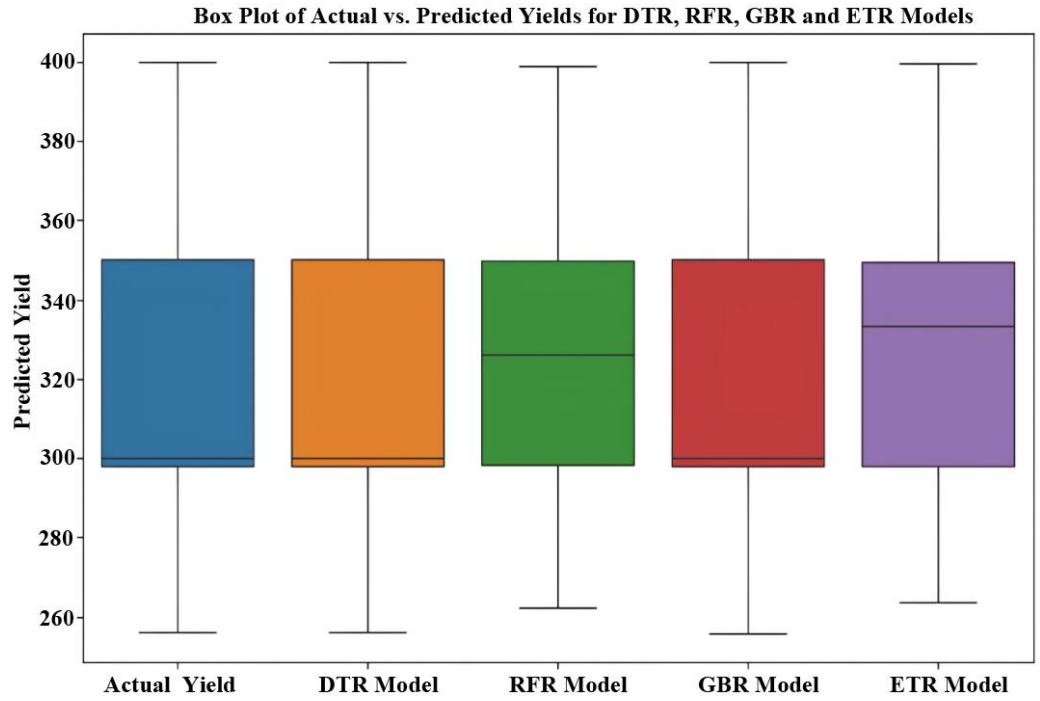

Figure 7. Anticipated coffee yield using the best four models for the best Group-2 characteristics and the actual measured coffee yield during testing

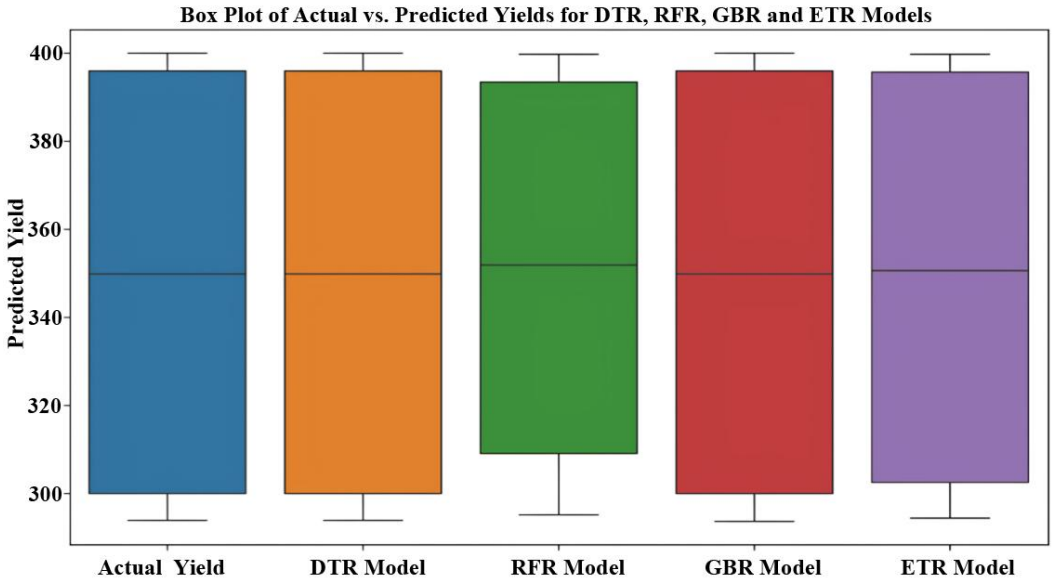

Figure 8. Anticipated coffee yield using the best four models for the best Group-3 characteristics and the actual measured coffee yield during testing

Table 19, Table 20, and Table 21 present the expected coffee yield data for each of the four stochastic models (ETR, GBR, RFR, and DTR) for the three optimal input combinations (Groups 1, 2, and 3) in terms of the model error variance in maximum, minimum, skewness, kurtosis, standard deviation, p25, p50 (median), and p75. Figure 6, Figure 7, and Figure 8 describe the overall estimated yield during the optimal testing phase. Furthermore, the models under consideration indicate only minor differences between the top and lower quartiles of the observed data, even though the higher quartile seemed to be under-predicted across all four models and all three sets of input parameters. The lowest quartiles of the models' predicted and observed data were almost the same.

Similar conclusions were derived using the minimum, standard deviation, and quartiles. Based on these performance errors, two deductions are possible:

(i). A comprehensive and robust statistical dependency analysis of inputs and the target variable must choose the most essential input variables to forecast coffee crop production based on agro-ecological factors (Table 1);

(ii). ETR models outperformed with an R^2 of 0.98 for C×R, an R^2 of 0.98 for Sln3, and an R^2 of 0.97 for Sln5B compared to GBR (an R^2 of 0.96 for C×R, an R^2 of 0.96 for Sln3, and an R^2 of 0.94 for Sln5B), RFR (an R^2 of 0.91 for C×R, an R^2 of 0.95 for Sln3, and an R^2 of 0.93 for Sln5B), and DTR (an R^2 of 0.91 for C×R, an R^2 of 0.90 for Sln3, and an R^2 of 1.0 for Sln5B) in forecasting coffee yields and identifying abiotic factor correlations.

4. Discussion

Several factors affect coffee production, including geographical area, soil fertility, plant age, rainfall amount, temperature, sunlight, humidity, vapor, dew points, and relative humidity. These factors impact the coffee harvest every year. It is not easy to manage manufacturing in a way that meets both supply and demand. Based on agro-ecological factor data from 2015–2022, the current study used four stochastic models to forecast the production of three coffee types: C×R, Sln3, and Sln5B. The results may help direct studies on the decision-making process in smallholder coffee farms, even if uncontrolled environmental variables affect coffee development. Other factors that may impact coffee production, such as pests and using fertilizers, were not discussed in this study. These characteristics could improve the accuracy of coffee production forecast models. This research used four stochastic models, such as ETR, RFR, GBR, and DTR, to predict coffee crop output using 2015–2022 agro-ecological factor data. Among the four stochastic machine learning models, ETR performed well with an accuracy of an R^2 of 0.98 for C×R, an R^2 of 0.98 for Sln3, and an R^2 of 0.97 for SLn5B coffee types in comparison with GBR (an R^2 of 0.96 for C×R, an R^2 of 0.96 for Sln3, and an R^2 of 0.94 for Sln5B), RFR (an R^2 of 0.91 for C×R, an R^2 of 0.95 for Sln3, and an R^2 of 0.93 for Sln5B), and DTR (an R^2 of 0.91 for C×R, an R^2 of 0.90 for Sln3, and an R^2 of 1.0 for Sln5B) by considering the three sets of optimal input parameters (Groups 1, 2, and 3) for 70:30 split ratios.

Machine learning has several practical uses in agriculture, such as predicting pest infestations (Kim et al., 2014), modeling crop growth and yield (Drummond et al., 2003; Fieuzal et al., 2017; Görgens et al., 2015; Kaul et al., 2005), predicting changes in groundwater levels (Sahoo et al., 2017), developing better irrigation techniques, practicing precision farming, or even mapping farmland (Chemura & Mutanga, 2016; Cheviron et al., 2016; Dimitriadis & Goumopoulos, 2008). Because of the automation of the variable selection procedure, the biological assumptions may be overlooked in the predictors selected. It was found in this study that the ETR model, with rainfall, age, and CLR occurrence included as predictors, produced the best results, highlighting the greatest impact of these variables on coffee output. Research on the impact of these variables on coffee development and bean production supports this finding. Soil factors have been identified in the literature as the most significant yield confounders. The results thus far demonstrate the validity of Table 7, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13, Table 14, Table 15, Table 16, Table 17, Table 18 inside the model developed for the three parameter sets (Figure 2, Figure 3, Figure 4, Figure 5), where the R^2 value compares the model's fit to a horizontal straight line. Existing research has not yet identified cases where the chosen model provides a poorer fit to the data than a straight line. There's room for improvement in the ETR model.

For example, a UK-based study by Pantazi et al. (2016) indicated that ANNs could accurately forecast wheat (*Triticum aestivum*) with an average overall harvest prediction accuracy of 78–82% and divide the land into zones with varying yield potential using multi-layer soil data and remotely sensed metrics of crop development. The findings of this current study reveal that the ETR model outperformed with an R^2 of 0.98 and an RMSE of 8.61 kg ha^{-1} of Group-1 parameters for C×R, an R^2 of 0.98 and an RMSE of 8.27 kg ha^{-1} of Group-3 parameters for Sln3, an R^2 of 0.97 and an RMSE of 7.79 kg ha^{-1} of Group-3 parameters for Sln5B coffee types compared to the GBR, RFR, and DTR models. The superior performance of ETR can be primarily attributed to its higher degree of randomization, its ability to manage multicollinearity in high-dimensional datasets, and its effective balance between bias and variance. Unlike RFR, which searches for the best split at each node, ETR selects split thresholds arbitrarily, introducing additional randomness into the model. This randomness enhances robustness against noise, a crucial factor in agro-ecological datasets characterized by strong correlations and nonlinear relationships. As a result, ETR consistently achieved higher R^2 (0.97–0.98) and lower RMSE (as low as 7.79 kg ha^{-1}) values compared to RFR and GBR.

Another important advantage of ETR lies in its ability to handle multiple interdependent agro-ecological predictors such as cloudiness (CLR), minimum and maximum temperature (Tmin and Tmax), relative humidity, rainfall, and soil nutrient parameters. These variables are often highly correlated, which can bias conventional ensemble models. By averaging over a large ensemble of randomized trees, ETR effectively reduces the dominance of strongly correlated features, thereby improving predictive stability. This strength was particularly evident in Group-3 datasets, which included a greater number of input parameters, where ETR delivered notably higher predictive accuracy. ETR also demonstrated a more favorable bias–variance trade-off compared to the other models. RFR was prone to higher variance, with performance deteriorating at larger training–testing splits (e.g., 50:50), whereas GBR was more sensitive to hyperparameters such as learning rate and sequential error correction, leading to bias and error accumulation. In contrast, ETR maintained robustness across varying split ratios (70:30 and 80:20), highlighting its resilience to dataset partitioning.

In addition to predictive accuracy, ETR offers computational simplicity and stability. Its randomized splitting

mechanism reduces computational overhead relative to GBR, which builds trees sequentially and demands greater processing resources. The ability of ETR to maintain stable performance across split ratios ranging from 90:10 to 70:30 further demonstrates its superior generalization to unseen datasets. Finally, scatterplot analyses and error diagnostics provided further evidence of ETR's effectiveness. Predicted versus observed yield values exhibited tighter clustering for ETR compared to GBR and RFR, indicating reduced error dispersion. Moreover, RMSE values for ETR were consistently 2–5 kg ha^{-1} lower than those of GBR and 6–10 kg ha^{-1} lower than those of RFR, confirming its enhanced predictive accuracy and reliability in coffee yield forecasting.

Although the ETR model performed better, the study has several limitations that require more testing. By carefully selecting the most relevant agro-ecological component for yield estimation, the ETR model could provide the groundwork for biophysical modeling of coffee crop production at larger geographical scales, drawing on data from varied parts of India that produce coffee. When developing the ETR model, it is necessary to include small- and large-scale consumption situations. Because both the farm and the dataset are subject to change, a random sampling method that generates an ensemble ETR model could amplify the constraints of statistical error and provide light on the uncertainty of predictions. Understanding the intricate dynamics of the environment, soil, pests, diseases, and agricultural techniques is crucial for modeling crop development. The Indian coffee business faces issues such as acid soils, intensive farming, and falling soil fertility. Improving crop nutrition and resource usage efficiency are crucial for sustainable production in the future. Key agro-ecological characteristics for yield estimation were discovered in this study using the proposed technique. While non-selected nutrients may still impact coffee development, the results of this study may inform future research on improving fertilization decisions in smallholder farms.

The study proposes improving prediction models using machine learning. The potential of information technology in coffee bean production can be evaluated broadly in two categories: (i) as a tool for direct contribution to coffee bean productivity; and (ii) as an indirect tool for empowering coffee planters to make informed and quality decisions, which will lead to better, consistent coffee yield in all adverse situations.

5. Conclusions

Optimal coffee crop production was evaluated at a coffee research station in Balehonnur, Karnataka, using stochastic machine learning models as a strong data-driven approach to analyzing predictive traits in agro-ecological factors data. Among these four models (ETR, GBR, RFR, and DTR) employed in the current work, the ETR model stands out with an accuracy of an R^2 of 0.98 and an RMSE of 8.61 kg ha^{-1} using Group-2 parameters for C×R, an R^2 of 0.98 and an RMSE of 8.27 kg ha^{-1} using Group-3 parameters for Sln3, and an R^2 of 0.97 and an RMSE of 7.79 kg ha^{-1} using Group-3 parameters for Sln5B coffee types in coffee-growing regions in India. To estimate the objective variable, coffee crop yield (*Y*), these models were used to examine various agro-ecological parameters, including agronomic, abiotic, and biotic elements. The results show that in the prediction of coffee yield using multiple inputs, ETR is more reliable and efficient at extracting features between agro-ecological factors and crop yields than GBR (an R^2 of 0.96 and an RMSE of 10.99 kg ha^{-1} using Group-3 parameters for C×R, an R^2 of 0.96 and an RMSE of 11.23 kg ha^{-1} using Group-3 parameters for Sln3, an R^2 of 0.94 and an RMSE of 11.67 kg ha^{-1} using Group-2 parameters for Sln5B), RFR (an R^2 of 0.91 and an RMSE of 14.12 kg ha^{-1} using Group-2 parameters for C×R, an R^2 of 0.95 and an RMSE of 10.71 kg ha^{-1} using Group-2 parameters for Sln3, an R^2 of 0.93 and an RMSE of 11.01 kg ha^{-1} using Group-3 parameters for Sln5B), and DTR (an R^2 of 0.91 and an RMSE of 16.10 kg ha^{-1} using Group-3 parameters for C×R, an R^2 of 0.90 and an RMSE of 15.10 kg ha^{-1} using Group-3 parameters for Sln3, an R^2 of 1.0 and an RMSE of 1.27 kg ha^{-1} using Group-3 parameters for Sln5B). Using a collection of meticulously curated datasets for agro-ecological factors, the current study validated the possible value of integrating AI algorithms with biophysical crop models in decision-support systems that employ precision agriculture. While the present study demonstrates the robustness of stochastic machine learning models, particularly ETR, in predicting coffee yields with high accuracy, certain limitations remain. The reliance on research station data may restrict generalizability to heterogeneous farmer-managed systems, highlighting the need for broader, multi-regional datasets. Future work should therefore focus on integrating diverse agro-ecological data sources, including remote sensing and real-time climatic variables, to enhance scalability and adaptability of the models. Refinement through hybrid approaches that combine AI algorithms with biophysical crop models could further strengthen decision-support frameworks for precision agriculture and sustainable coffee production.

Author Contributions

Conceptualization, C.S.S. and K.K.U.; methodology, C.S.S.; software, C.S.S.; validation, C.S.S., K.K.U., and K.N.; formal analysis, C.S.S.; investigation, C.S.S.; resources, K.K.U.; data curation, C.S.S., V.H., and K.N.; writing—original draft preparation, C.S.S.; writing—review and editing, K.N. and K.K.U.; visualization, K.N.; supervision, K.K.U.; project administration, K.K.U.; funding acquisition, K.K.U. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgement

We gratefully acknowledge the historical coffee records provided by the Central Coffee Research Institute (CCRI), Balehonnur, Chikkamagaluru District, Karnataka, India, through JSS Science and Technology University, Mysuru. These records were instrumental in forecasting coffee harvest yields and in enhancing our understanding of the diverse factors influencing coffee production.

Conflict of Interest

The authors declare no conflict of interest.

References

Aceves Navarro, L. A., Rivera Hernández, B., López Castañeda, A., Palma López, D. J., González Mancillas, R., & Juárez López, J. F. (2018). Potential areas and vulnerability of the robust coffee crop (*Coffea canephora* P.) to climate change in the state of Tabasco, Mexico. *Nova Sci., 10* (20), 369–396. https://doi.org/10.21640/ns.v10i20.1379.

Campanha, M. M., Santos, R. H. S., de Freitas, G. B., Martinez, H. E. P., Garcia, S. L. R., & Finger, F. L. (2004). Growth and yield of coffee plants in agroforestry and monoculture systems in Minas Gerais, Brazil. *Agroforest Syst., 63*, 75–82. https://doi.org/10.1023/B:AGFO.0000049435.22512.2d.

Cerda, R., Avelino, J., Gary, C., Tixier, P., Lechevallier, E., & Allinne, C. (2017). Primary and secondary yield losses caused by pests and diseases: Assessment and modeling in coffee. *PloS One*, *12*(1), e0169133. https://doi.org/10.1371/journal.pone.0169133.

Chai, T. & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chemura, A. & Mutanga, O. (2016). Developing detailed age-specific thematic maps for coffee (*Coffea arabica* L.) in heterogeneous agricultural landscapes using random forests applied on Landsat 8 multispectral sensor. *Geocarto Int., 32*(7), 759–776. https://doi.org/10.1080/10106049.2016.1178812.

Cheviron, B., Vervoort, R. W., Albasha, R., Dairon, R., Le Priol, C., & Mailhol, J. C. (2016). A framework to use crop models for multi-objective constrained optimization of irrigation strategies. *Environ. Model. Softw., 86*, 145–157. https://doi.org/10.1016/j.envsoft.2016.09.001.

Chowdhury, D., Sarkar, M., Haider, M. Z., & Alam, T. (2018). Zone wise hourly load prediction using regression decision tree model. In *2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh* (pp. 1–6). https://doi.org/10.1109/CIET.2018.8660781.

Dimitriadis, S. & Goumopoulos, C. (2008). Applying machine learning to extract new knowledge in precision agriculture applications. In *2008 Panhellenic Conference on Informatics, Samos, Greece* (pp. 100–104). https://doi.org/10.1109/PCI.2008.30.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., & Kitchen, N. R. (2003). Statistical and neural methods for site–specific yield prediction. *Trans. ASAE*, *46*(1), 5–14. https://doi.org/10.13031/2013.12541.

Fieuzal, R., Marais Sicre, C., & Baup, F. (2017). Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. *Int. J. Appl. Earth Obs. Geoinf., 57*, 14–23. https://doi.org/10.1016/j.jag.2016.12.011.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn., 63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Görgens, E. B., Montaghi, A., & Rodriguez, L. C. E. (2015). A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. *Comput. Electron. Agric., 116*, 221–227. https://doi.org/10.1016/j.compag.2015.07.004.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agric. Syst., 85*(1), 1–18. https://doi.org/10.1016/j.agsy.2004.07.009.

Kim, Y. H., Yoo, S. J., Gu, Y. H., Lim, J. H., Han, D., & Baik, S. W. (2014). Crop pests prediction method using regression and machine learning technology: Survey. *IERI Procedia, 6*, 52–56. https://doi.org/10.1016/j.ieri.2014.03.009.

Kittichotsatsawat, Y., Tippayawong, N., & Tippayawong, K. Y. (2022) Prediction of annual coffee production yield using artificial neural network and multiple linear regression techniques [Preprint]. *Research Square*. https://doi.org/10.21203/rs.3.rs-1504007/v1.

Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., Mushtaq, S., & Nguyen, V. P. (2018). Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Comp. Electron. Agric.*, *155*, 324–338. https://doi.org/10.1016/j.compag.2018.10.014.

Legates, D. R. & McCabe, G. J. (2013). A refined index of model performance: A rejoinder. *Int. J. Climatol.*, *33*(4), 1053–1056. https://doi.org/10.1002/joc.3487.

Legates, D. R. & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018.

Maro, G. P., Mrema, J. P. P., Msanya, B. M., Janssen, B. H. H., & Teri, J. M. (2014). Developing a coffee yield prediction and integrated soil fertility management recommendation model for Northern Tanzania. *Int. J. Plant Soil Sci.*, *3*(4), 380–396. https://doi.org/10.9734/IJPSS/2014/6883.

Muliasari, A. A. & Dewi, H. (2022). Estimated yield potential of robusta coffee (*Coffea canephora* Pierre ex A. Froehner) at Bogor District. *E3S Web Conf.*, *348*, 00020. https://doi.org/10.1051/e3sconf/202234800020.

Natarajan, R., Subramanian, J., & Papageorgiou, E. I. (2016). Hybrid learning of fuzzy cognitive maps for sugarcane yield classification. *Comput. Electron. Agric.*, *127*, 147–157. https://doi.org/10.1016/j.compag.2016.05.016.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Comput. Electron. Agric.*, *121*, 57–65. https://doi.org/10.1016/j.compag.2015.11.018.

Romero-Alvarado, Y., Soto-Pinto, L., García-Barrios, L., & Barrera-Gaytán, J. F. (2002). Coffee yields and soil nutrients under the shades of *Inga* sp. vs. multiple species in Chiapas, Mexico. *Agrofor. Syst.*, *54*, 215–224. https://doi.org/10.1023/A:1016013730154.

Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resour. Res.*, *53*(5), 3878–3895. https://doi.org/10.1002/2016WR019933.

Santhosh, C. S. & Umesh, K. K. (2022). A compendium probabilistic prospective for predicting coffee crop yield based on agronomical factors. In *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India*, (pp. 1–8). https://doi.org/10.1109/ICERECT56837.2022.10060420.

Santhosh, C. S. & Umesh, K. K. (2023). An ensemble approach for coffee crop yield prediction based on agronomic factors. *ASEAN Eng. J.*, *13*(3), 29–38. https://doi.org/10.11113/aej.v13.18846.

Santhosh, C. S. & Umesh, K. K. (2024). A method based on artificial intelligence that predicts arabica coffee yield by analysing abiotic factors and the prevalence of coffee leaf rust. *Int. Res. J. Adv. Eng. Manag.*, *2*(5), 1450–1465. https://doi.org/10.47392/IRJAEM.2024.0196.

Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). Agricultural production output prediction using supervised machine learning techniques. In *2017 1st International Conference on Next Generation Computing Applications (NextComp), Mauritius* (pp. 182–187). https://doi.org/10.1109/NEXTCOMP.2017.8016196.

Shastry, K. A., Sanjay, H. A., & Deshmukh, A. (2016). A parameter based customized artificial neural network model for crop yield prediction. *J. Artif. Intell.*, *9*, 23–32. https://doi.org/10.3923/jai.2016.23.32.

Singh, U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies*, *14*(16), 5196. https://doi.org/10.3390/en14165196.

Singh, V., Sarwar, A., & Sharma, V. (2017). Analysis of soil and prediction of crop yield (rice) using machine learning approach. *Int. J. Adv. Res. Comput. Sci.*, *8*(5), 1254.

Sudha, M., Machenahalli, S., Giri, M. S., Ranjini, A. P., & Daivasikamani, S. (2020). Influence of abiotic factors on coffee leaf rust disease caused by the fungus *Hemileia vastatrix* Berk. & Br. under changing climate. *J. Agrometeorol.*, *22*(3), 367–371.

Tadesse, Y., Amare, D., & Kesho, A. (2021). Coffee leaf rust disease and climate change. *World J. Agric. Sci.*, *17*(5), 418–429. https://doi.org/10.5829/idosi.wjas.2021.418.429.

Varshitha, D. N. & Choudhary, S. (2022). Soil fertility and yield prediction of coffee plantation using machine learning technique. *Res. J. Agric. Sci.*, *13*(2), 514–518.

Wang, N., Jassogne, L., van Asten, P. J. A., Mukasa, D., Wanyama, I., Kagezi, G., & Giller, K. E. (2015). Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. *Europ. J. Agron.*, *63*, 1–11. https://doi.org/10.1016/j.eja.2014.11.003.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *Int. J. Climatol.*, *32*(13), 2088–2094. https://doi.org/10.1002/joc.2419.