Smart Agricultural Technology 14 (2026) 102256

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Arabica coffee fruit load non-destructive estimation using machine learning techniques and yield simulation

Luis Carlos Imbachí-Quinchua [a], Carlos Andrés Unigarro [b], Álvaro Gaitán-Bustamente [c], Andrés Felipe León-Burgos [d,*]

[a] Department of Biometrics, National Coffee Research Center-Cenicafé. Manizales 170009, Colombia
[b] Department of Plant Physiology, National Coffee Research Center-Cenicafé. Manizales 170009, Colombia
[c] Technical Manager, National Federation of Coffee Growers of Colombia, Bogotá 111321, Colombia
[d] Department of Crop Science, National Coffee Research Center-Cenicafé. Manizales 170009, Colombia

ARTICLE INFO

Keywords:
Agricultural planning
Crop variability
Leaf area-to-fruit ratio
Monte Carlo method
Stochastic simulation

ABSTRACT

Methods for estimating coffee yield employ expensive and destructive sampling techniques that offer limited flexibility in describing agricultural data variability. The objective of this study was to investigate machine learning (ML) techniques to develop a model that predicts fruit load from nondestructive shoot vegetative growth measurements, integrated with a probabilistic approach for simulating crop yields. Evaluations were conducted on Castillo® Centro variety plants for four consecutive years from sowing to establishment. For ML modeling, data on foliar formation, plagiotropic branch growth, and yield components were collected over three productive years. Three ML techniques were evaluated to estimate fruit load: support vector regression (SVR), artificial neural network (ANN), and random forest (RF). Based on probabilistic distributions from 120 trees, a tree-level yield simulation was conducted, generating a simulated population of 1200 trees. The two most productive branches of each tree were used to parameterize the distributions, incorporating the residual components of the ML models directly into the simulation process. Yield was defined as production per tree in grams (g). The ANN model exhibited the best performance, explaining >95% of data variability ($R^2 = 0.98$) and the lowest dispersion (RMSE = 3.64 fruits branch^{-1}), with foliar formation contributing 84% to the model structure. The mean difference between simulated and observed yields during the first harvest year did not exceed 300 g per plant. These findings reveal that integrating ML methods with stochastic processes is a robust approach for coffee yield prediction and simulation.

1. Introduction

The estimation of coffee production is key for agricultural planning, the efficient use of resources and decision-making in terms of agronomic management for guaranteeing the success of the crop [1,2]. Currently, yields are calculated by quantifying flower buds during preanthesis and fruit collection using expensive methods that are difficult to implement in large areas, which limits their applicability [3,4]. Consequently, nondestructive, fast and easily implementable methods are needed that integrate advanced tools such as machine learning (ML), which have demonstrated higher accuracy in the estimation of crop yields than conventional statistical models [5,6]. These models facilitate the integration of agrometeorological, physiological and management variables, which enables holistic interpretability [7,8].

ML models such as random forest (RF), support vector machine (SVM), and neural networks (NNs) are widely adopted as key tools for improving coffee yield estimation [7,9]. In Brazil, ML approaches integrating temporal information and satellite imagery provide robust yield estimation and productivity mapping at the plot level, with RF showing strong predictive performance [8]. Previous studies also indicate that RF achieves high accuracy in coffee yield estimation when integrating cultivar, climate, and management-related variables [9]. Overall, these techniques demonstrate strong capability for identifying complex structures, capturing nonlinear relationships, and adapting to variable information patterns [10].

The statistical distributions such as gamma, Weibull, log-normal and even adapted variants have proven essential for improving crop prediction by capturing real patterns of variability and asymmetry in

* Corresponding author.
E-mail address: felipeleonb27@gmail.com (A.F. León-Burgos).

https://doi.org/10.1016/j.atech.2026.102256
Received 21 April 2026; Received in revised form 27 May 2026; Accepted 27 May 2026
Available online 30 May 2026

agricultural data [11,12].The gamma and Weibull distributions are widely used in probabilistic analyses due to their outstanding performance in representing agricultural yields with positive skewness, providing a robust approximation adapted to biological processes for continuous variables such as production and plant growth in crops such as *Zea mays* [12,13], While the log-normal distribution has been applied to model canopy development and growth-related variables associated with biomass in crops such as *Oryza sativa* [14]. Together, these distributions allow for precise adjustment of the observed data and generate more robust probabilistic forecasts for integration with crop modeling.

Colombia is among the world's main exporters of Arabica coffee (*Coffea arabica* L.) and is recognized for its diverse microclimates between 3° and 7° north latitude, with altitudinal ranges from 1000 to 2000 masl [15]. These geographical differences promote variations in climatic conditions and influence crop phenology, which affects the patterns of flowering and harvest as well as the intensity of the simultaneous growth of vegetative and reproductive organs [16,17]. Under cultivable conditions in Colombia, a correlation between coffee fruits, leaf formation and the plagiotropic growth of branches has been observed, and the nature of this association directly influences the fruit load per plant, the future patterns of plant growth, harvest and measures of yield components such as the number of fruits at the branch or plant level, individual fruit mass, and percentage of malformed fruits [18–20].

The ratio of leaf area per fruit is a useful growth parameter to explain the interaction between vegetative and reproductive growth with increasing fruit load in *C. arabica* [21,22]. This parameter is used because it has been determined that a coffee fruit requires 20 cm^2 of leaf area, and this measure is sensitive to increases in fruit load influenced by climatic conditions and the establishment of the crop under either full-sun exposure or under shade [18,23,24]. In fact, it has been implemented as an input for statistical models that predict the fruit load per plant of *C. arabica* and to elucidate the growth dynamics of the coffee fruit and its influence on cup quality [3,25].

Despite recent advances in machine learning (ML) for agricultural yield estimation and crop monitoring using non-destructive measurements, its integration with stochastic simulations in coffee production remains limited. Most studies have developed predictive models and simulation analyses separately, with little attention given to incorporating ML-derived residual variability and uncertainty into yield simulations. Therefore, integrated methodologies are still needed to better represent production variability and strengthen crop yield estimation based on non-destructive shoot measurements [7,26,27].

Therefore, the objective of this study was to explore different ML techniques to characterize and adjust the residual error component associated with estimating fruit load from aboveground plant measurements (non-destructive), and thereby construct inputs for simulating crop yields across different production years. For the ML model structure, plant data related to growth and fruit production were collected during three productive years from cultivated *C. arabica* L. variety Castillo® Centro plants. Yield simulations were then performed for the first and second productive years, which showed highly contrasting harvests in terms of fruit load per plant under Colombian growing conditions [28]. This information is necessary for crop management decision-making in terms of labor planning with respect to manual or assisted harvesting and phytosanitary control and to identify new statistical methods that complement the estimation of crop production in Colombia and for coffee-growing countries worldwide, based on non-destructive methodologies.

2. Materials and methods

2.1. Study area, plant material, and climate

This research was conducted under field conditions for four consecutive years with plants of the Castillo® Centro variety established at the Naranjal Experimental Station affiliated with the National Center for Coffee Research-Cenicafé, located in the municipality of Chinchiná, Department of Caldas (04° 58 ' N; 75° 39′ W), at an altitude of 1381 m. In general, the climatic conditions during the productive years (2022 to 2024) were as follows: An average daily temperature of 21.35°C, an average daily photosynthetically active radiation (PAR) of 383.72 µmol m^2 s^{-1}, an average cumulative annual sunlight of 1,573 h and cumulative average annual precipitation of 2,821 mm (Fig. 1). The climate data were obtained from the conventional and automatic meteorological networks of the National Federation of Coffee Growers of Colombia, available on the Agroclima platform [29].

The plantation was established under exposure to the sun and planted in March 2021 with a distance of 1.25 m between rows and 1.0 m between plants achieving a density of 8000 plants ha^{-1}. The plant is a multiline variety obtained from the crossing of Caturra with the Timor hybrid, is resistant to coffee rust (*Hemileia vastatrix*) and coffee cherry disease caused by *Colletotrichum kahawae*, is highly productive, has good physical characteristics of beans, and is grown in approximately 85% of the arable area for coffee in Colombia [30]. For planting, 30 cm long × 30 cm wide × 30 cm deep holes were made, and 100 g $plant^{-1}$ lime was applied. Agronomic practices for commercial production related to liming, fertilization, and integrated management of the coffee berry borer (*Hypothenemus hampei*), mealybugs (Pseudococcidae) and weeds were performed as recommended by Cenicafé [31].

2.2. Crop experimental data and measurements

For data collection in the field, 40 experimental plots having a total area of 2200 m^2 were established, each with an area of 52 m^2 and consisting of 30 effective plants. After 16 months of planting, crop data were collected from 2022 to 2024 at the beginning of the productive stage for plants cultivated under the conditions of the central coffee-growing region of Colombia [28]. The timepoints of measurement for each year were defined on the basis of the flowering events recorded according to the methodology of Rendón and Montoya [32] and by monitoring the phenological stages of coffee fruit development according to the BBCH scale (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie), following the classification described by Arcila-Pulgarín et al. [33].

2.2.1. Measures of shoot growth

Measurements were obtained for a total of 120 plants, and in each experimental plot, three plants were randomly selected and monitored for all years. Then, in each plant and per year, two plagiotropic branches with the largest number of fruits in the productive third of the canopy were selected. Each sampling was carried out on the months of June and July of each year (2022 to 2024), which accounted for 60 to 80% of the total fruits produced during the year, and corresponded to the main harvest in the study area, in accordance with previous reports by León-Burgos et al. [18] and Rendón [28]. The years 2022 and 2023 corresponded to the productive periods used for yield evaluation and model development, whereas 2024 included complementary vegetative measurements associated with the continuation of plant growth.

Growth measures such as branch length (BL), total number of nodes per branch (TNN), total number of leaves per branch (TNL), accumulated leaf area per branch (ALA), leaf area-to-fruit ratio (LAFR), total number of fruiting nodes per branch (TNFN), number of fruits per branch (FB) and individual fruit mass (FM) were evaluated. The TNN, TNL, TNFN and FB were estimated by direct counts in each of the branches. The BL was measured from the base where it joins the main stem to the apex of the branch using a flexometer.

To calculate the accumulated leaf area per branch (ALA), Eq. (1) proposed by Unigarro et al. [34] was used to estimate the leaf area of each individual leaf (ELA). For this purpose, the length of the leaf blade (excluding the petiole) and the width of the middle part of each leaf were measured using a ruler. Subsequently, the estimated leaf area values obtained for each leaf were summed to determine the total

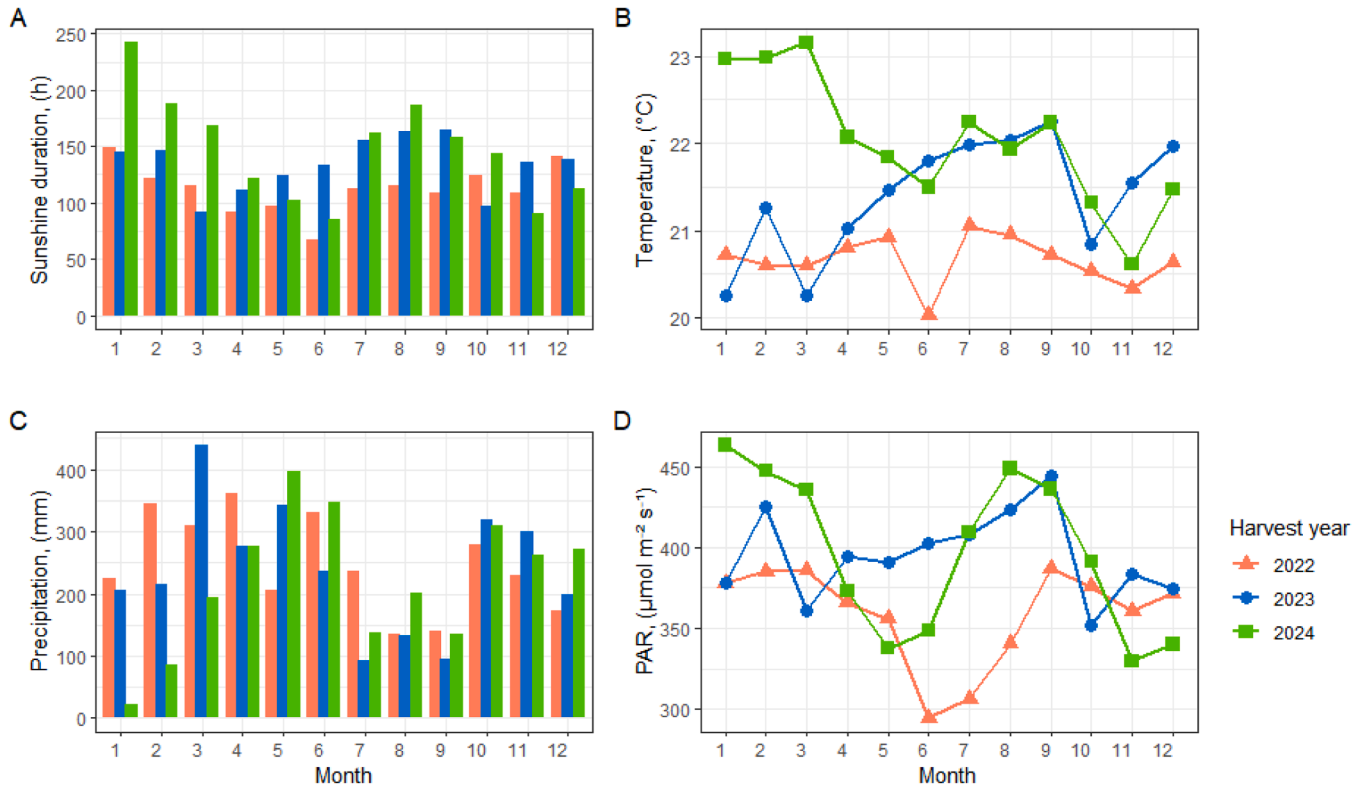

Fig. 1. Monthly climatic data for each year of production. (A) Accumulated solar brightness, (B) average temperature, (C) accumulated precipitation and (D) photosynthetically active radiation (PAR).

accumulated leaf area per branch (ALA). In parallel, the leaf area-to-fruit ratio (LAFR) was calculated as LAFR = ALA/FB, as reported by León-Burgos et al. [19].

$$ELA = 0.99927 * (L * (-0.14757 + 0.60986 * W)) \quad (1)$$

where ELA is the estimation of the leaf area, L is the leaf length and W is the leaf width.

Finally, for measurement of the FM, 100 fruits were collected from the total harvest of each experimental plot and weighed using an analytical balance with a precision of 0.01 g. Completely mature fruits that did not have physical or phytosanitary damage were selected, and the measurements were made on the main harvest of the study area according to the description by León-Burgos et al. [19]. A total of 4000 coffee cherries fruits were evaluated for each production year.

2.2.2. Crop yield

Fully ripe red fruits classified in the BBCH88 phenological stage were manually harvested, and the production of coffee cherries per plot in kg was recorded using a dynamometer with a precision of 100 g. The harvest frequency was every 18 days from August 2022 to December 2024.

2.3. Agricultural modeling for fruit load estimation and coffee yield simulation

2.3.1. Optimization and performance of machine learning techniques for determining fruit load

All the growth measurements recorded during the three production years (2022 to 2024) were used as input (BL, TNN, TNL, ALA, LAFR, TNFN), and three ML models were evaluated to predict the fruit load by measuring the FB in relation to shoot growth: support vector regression, artificial neural network and random forest. Model performance was assessed using a k-fold cross-validation scheme, in which the dataset was partitioned into k subsets, iteratively using one fold for validation and the remaining folds for training. A nested cross-validation approach was implemented, with an inner k-fold procedure for hyperparameter tuning and an outer k-fold procedure for model evaluation, ensuring unbiased performance estimation. For the SVR model, hyperparameters (C and γ) were optimized using a radial kernel within the inner cross-validation loop. The RF model was configured with 1000 trees, and the number of variables randomly selected at each split was optimized within each training fold. For the ANN, a multilayer architecture with six hidden neurons was implemented and trained using backpropagation. Prior to training, predictor variables were normalized (2) within each training fold, and the same scaling parameters were applied to the corresponding validation fold to prevent data leakage [35]. Model performance for each algorithm (SVR, RF, and ANN) was quantified using out-of-sample predictions obtained from the outer cross-validation, and metrics including RMSE and R^2 were calculated [36].

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \quad (2)$$

where, Zij: the standardized value (z-score) of individual *i* for variable *j*, X_{ij} is the value of individual *i* in variable *j* and μ_j and σ_j are the mean and standard deviation of variable *j*, respectively.

2.3.2. Incorporation of uncertainty in fruit load estimation

Once the ML model was defined to estimate the FB, a *skewed-normal* distribution was fitted to the residuals (D_{res}) obtained in the validation stage to characterize the variability not explained by the model and complement the deterministic estimation of the fruit load. This procedure allows the identification of possible asymmetries in the distribution of errors associated with the nature of the crop and the influence of unobserved factors. The adjusted distribution was subsequently used to alter the empirical fruit counts (observed in the field) by adding a random term attached to this distribution, which strengthens the ability

of the approach to represent realistic production scenarios. In this way, the final predictions reflect not only the expected value of the fruit load but also the uncertainty inherent to the production process, providing a more robust basis for analysis and agronomic decision-making.

2.3.3. Parameter calibration and crop yield simulation

Simulation of coffee cherry yield was carried out for the first and second productive years per tree; these harvests presented highly contrasting fruit loads in Colombia's coffee-growing region [20]. The relative contribution of the number of fruits on the two productive branches evaluated to the total number of fruits per tree was subsequently determined by considering the number of fruits per branch and the total number of productive branches per tree, and the estimation was performed using Eq. (3).

$$CPBT_i = \frac{\sum_{j=1}^{2} FB_{ij}}{\sum_{k=1}^{PB_i} FB_{ik}} \tag{3}$$

where $CPBT_i$ is the contribution of the two productive branches in tree i, BF_{ij} is the number of fruits on branch j of tree i ($j = 1,2$ for the most productive branches), PB_i is the total number of productive branches on tree i, and FB_{ik} is the number of fruits in branch k of tree i, with k = 1, 2, …, PB_i.

For the productive years according to the guide for the graph by Cullen and Frey [37], the empirical distributions were fitted to the number of fruits predicted by the ML model (FB_{ml}), $CPBT_i$ and FM. In accordance with the above, the gamma and beta distributions were analyzed for $CPBT_i$, and the gamma, Weibull and normal distributions were analyzed for the FM and FB_{ml} (Fig. S1). The selection of the type of distribution was based on the AIC and BIC information criteria, and its goodness of fit was verified by the Anderson–Darling test with a significance level of 5%.

Once the probability distributions mentioned above were adjusted, a stochastic simulation was performed using the Monte Carlo approach [38] adapted for the 40 experimental plots, each consisting of 30 trees with two productive branches per tree. Under this structure, for each productive year, the two productive branches were simulated, and in each branch, the number of fruits was assigned from the FB_{ml} adjusted by D_{res}, where a residual component obtained from the distribution of residuals of the ML model was added to the distribution of predicted fruits to generate a more objective estimate of the number of fruits per branch. Each simulated fruit was subsequently assigned a mass according to the FM distribution, using which the production per branch was estimated. Finally, with $CPBT_i$, the simulated values were integrated to obtain the total production per tree. However, the previous estimation reflects potential productive behavior; therefore, a correction factor described in Eq. (4) was added to adjust the simulated production per tree in each productive year:

$$\underline{CF} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\frac{P_i}{\underline{R}}}{\frac{1}{n}\sum_{j=1}^{n} N_{ij}}\right) \tag{4}$$

where CF is the correction factor, P_i corresponds to the total production of tree i, $\underline{R}$ is the average number of branches per tree, N_{ij} is the number of fruits per branch j of tree i, and m and n are the number of trees and the number of branches sampled, respectively.

Finally, for each productive year, to stabilize the variance and evaluate the uncertainty associated with the average production per tree observed, a nonparametric bootstrap procedure was applied with 1200 replicates of equal size, obtained by random sampling with replacement [39]. All analyses were performed in R software version 4.2.1 [40]. The following R packages were used for the analyses: *fitdistrplus* (by adjusted for distribution), *corrplot* (for correlation analysis and visualization), *randomForest* (for random forest modeling), *nnet* (for neural network modeling), *NeuralNetTools* (for neural network interpretation and visualization), *e1071* (for support vector machine modeling), and *ggplot2* (for data visualization).

3. Results

3.1. Machine learning model fitting for fruit load estimation with residual error correction

The performance of the ML techniques in predicting fruit load was shown in Fig. 3. The results of the cross-validation of the SVR model revealed an adequate fit between the observed and predicted values of the number of fruits per branch. The number of fruits was concentrated and distributed close to the 45° reference line in the entire range of the observed values, which indicated agreement between the predicted and actual measurement, without indications of systematic bias. The coefficient of determination (R^2) obtained was 0.94, indicating that the model explained approximately 94% of the total variability, with an average prediction error (RMSE) of 9.56 fruits per branch, corresponding to approximately 7.1% of the mean observed fruits per branch (134 fruits) (Fig. 2A).

The combination of the R^2 criteria close to 1.0 and a reduced mean error suggested that the SVR model was valid for predictive purposes. However, when the performance of the ANN and SVR models was compared with respect to that of the RF (Fig. 2c), the ANN and SVR models exhibited greater precision and less dispersion of the number of fruits per branch predicted with the fit metrics, since an R^2 above 95% was obtained and the mean error of fruits per branch ranged from 3.64 to 9.56 (Fig. 2a and 2b). Although differences among these structures were evident, they reflected underlying patterns that were difficult to capture with simple regression approaches, given that such models did not achieve an R^2 above 65% and exhibited an RMSE of 23.67 fruits $branch^{-1}$, corresponding to approximately 17.6% of the mean observed fruits per branch (Fig. S4).

The results indicated that the ANN was the most appropriate method for estimating the number of fruits per branch based on aboveground plant growth measurements in coffee plants. From this predictive performance, the probabilistic characterization of the errors of the model played a fundamental role, since it allowed the deterministic estimation to be extended toward an approach that incorporated the uncertainty inherent to the process. Consequently, the disturbance of the observed fruit counts by simulated residues from the fitted distribution allowed the uncertainty associated with the prediction of the ANN model and the field count method to be integrated. This procedure did not assume that the observed counts were free of error but did recognize the presence of random variability and possible inaccuracies inherent to the measurement. Thus, incorporating residues helped compensate for random errors and more realistically represented the expected variability in the fruit load.

Additionally, the LARF and ALA measures constituted approximately 84% of the total contribution of the ANN model (Fig. S2), and part of the residual variability can be interpreted as the propagation of the uncertainty associated with these dominant variables in the estimation process.

3.2. Integration of parameter standardization and machine learning for Arabica coffee yield simulation

The measures $CPBT_i$, FB_{ml} and FM were standardized via the probabilistic approach for the simulation of coffee cherry yield per tree in this study (Table 1). The ANN model was also integrated to determine the distributions of FB_{ml}, which presented the best fit according to the statistical criteria described in the previous section. In this way, the null hypothesis for $CPBT_i$, was rejected based on the goodness-of-fit analysis, confirming that both gamma and beta distributions provided an adequate representation of the data. However, the beta distribution was

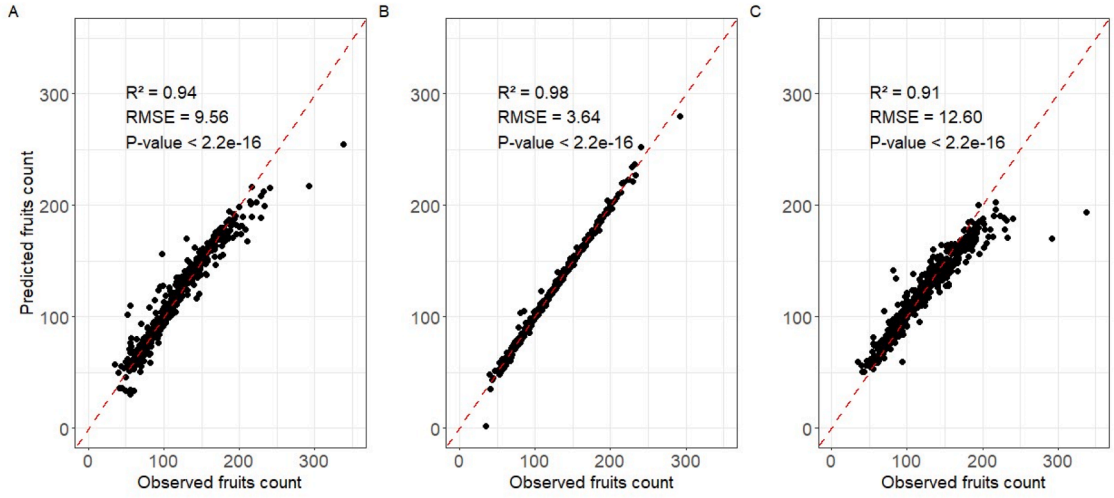

Fig. 2. Results of cross-validation of the ML techniques evaluated for fruit load prediction via the number of fruits per branch in the three productive years. (A) Support vector regression (SVR), (B) artificial neural network (ANN) and (C) random forest (RF).

Table 1
Estimation of parameters and the statistical criteria for adjusting the probabilistic distributions in the contribution of the productive branches evaluated in the tree ($CPBT_i$), the number of fruits per branch predicted by ML(FB_{ml}) and the individual fruit mass(FM) for the first and second years of crop production.

Measure	Distribution	Parameters	AIC	BIC	Anderson−Darling Value	*P* value
$CPBT_i$	Gamma	α =31.72 λ=195.92	−509.83	−504.25	0.092073	0.261
	Beta	α = 26.04 β=134.79	−507.23	−501.66	0.09638	0.214
FB_{ml}	Gamma	α =9.25 λ=0.081	5742.15	5750.85	0.97083	0.3729
	Weibull	λ = 3.23 k = 127.39	5785.03	5793.70	3.3804	0.01763
	Normal	μ = 114.18 σ=37.54	5780.23	5788.93	3.2943	0.001946
FM	Gamma	α =41.33 λ=18.97	3198.76	3211.74	2.712	0.05
	Weibull	λ = 6.63 k = 2.32	3730.53	3743.5	5.9571	0.001
	Normal	μ = 2.17 σ=0.33	3236.27	3249.25	46.534	1.23E-07

selected because it showed the lowest AIC and BIC values among the evaluated distributions, indicating the best overall fit. Similarly, the fit of each of the evaluated distributions is shown in Fig. 4a, in which the range of $CPBT_i$ varies from 10% to 26%, but with the beta distribution fit, the mean is equivalent to 16.1% of the total fruit production on the tree.

For the adjustment of the distributions for FBml and FM, the gamma distribution had the lowest AIC and BIC values, and according to the p value of the goodness of fit test, the null hypothesis was not rejected, which indicated that under this distribution, the behavior for these measures was adequately and accurately represented (Table 1). The mean values for FB_{ml} (114) and FM (2.17 g) recorded in the first and second years of production of the Castillo® Centro cultivar are reported in Figs. 3b and 3c.

Once the probability distributions for $CPBT_i$, FB_{ml} and FM were defined, we proceeded to simulate the coffee cherry production first at the branch level and then scaled it to the tree level for the first year of production (Fig. 4). The distribution of the simulated production ranged from 735 g to 7000 g, with an average of 3251 g. However, this did not reflect the average observed production per tree (1171 g), which corroborated the need to apply a correction factor to reduce simulated production values and adjust to the observed values (Fig. 4b). The values of the correction factor estimated in this study were shown in Figure S3, which fluctuated between 0.26 and 0.70 for the first year of production and between 0.46 and 0.98 for the second year of production, and were used as reference values for the simulation.

A nonparametric bootstrap resampling procedure was implemented from the observed production to acquire a more solid statistical basis for interpretation of the structure of the simulated yields (Fig. 5). In the first year, the average observed production was 1171 g per tree (Fig. 5A), while the simulated production reached 1462 g (Fig. 5B). For the second year, the average observed value was 3926 g per tree (Fig. 5C) compared with 3156 g obtained in the simulation (Fig. 5D), an overestimation in the first year (291 g per tree) and an underestimation in the second (770 g per tree). In general, the simulation constitutes a robust and functional tool to represent the productive dynamics of each harvest. In addition, this bias can be corrected by adjusting the parameters used in the simulation, improving the accuracy of the model.

4. Discussion

In our research, different ML techniques were evaluated with the specific purpose of characterizing and adjusting the error component associated with the estimation of fruit load from non-destructive measurements of leaf formation, branch growth and yields components. An adjusted model was used to describe the structure of the residual variability and generate stochastic disturbances in the observed counts, thus allowing a more realistic simulation of crop yields in different production years, without the assumption of error-free measurements. These results constitute essential tools for optimizing agronomic crop management, focused on anticipating fruit availability, efficiently planning key processes such as labor organization, harvest scheduling, and decision-making associated with crop management practices [4,28]. In this way, the integration of advanced statistical techniques such as ML with non-destructive sampling techniques and a probabilistic approach represents a robust alternative for the simulation of coffee crop yield adapted to the biological processes of plants [10]

ANN exhibited higher performance than RF and SVR models for FB estimation (Fig. 2). This model has previously been applied for predicting C. arabica cherries yield in arable regions of Thailand [41].

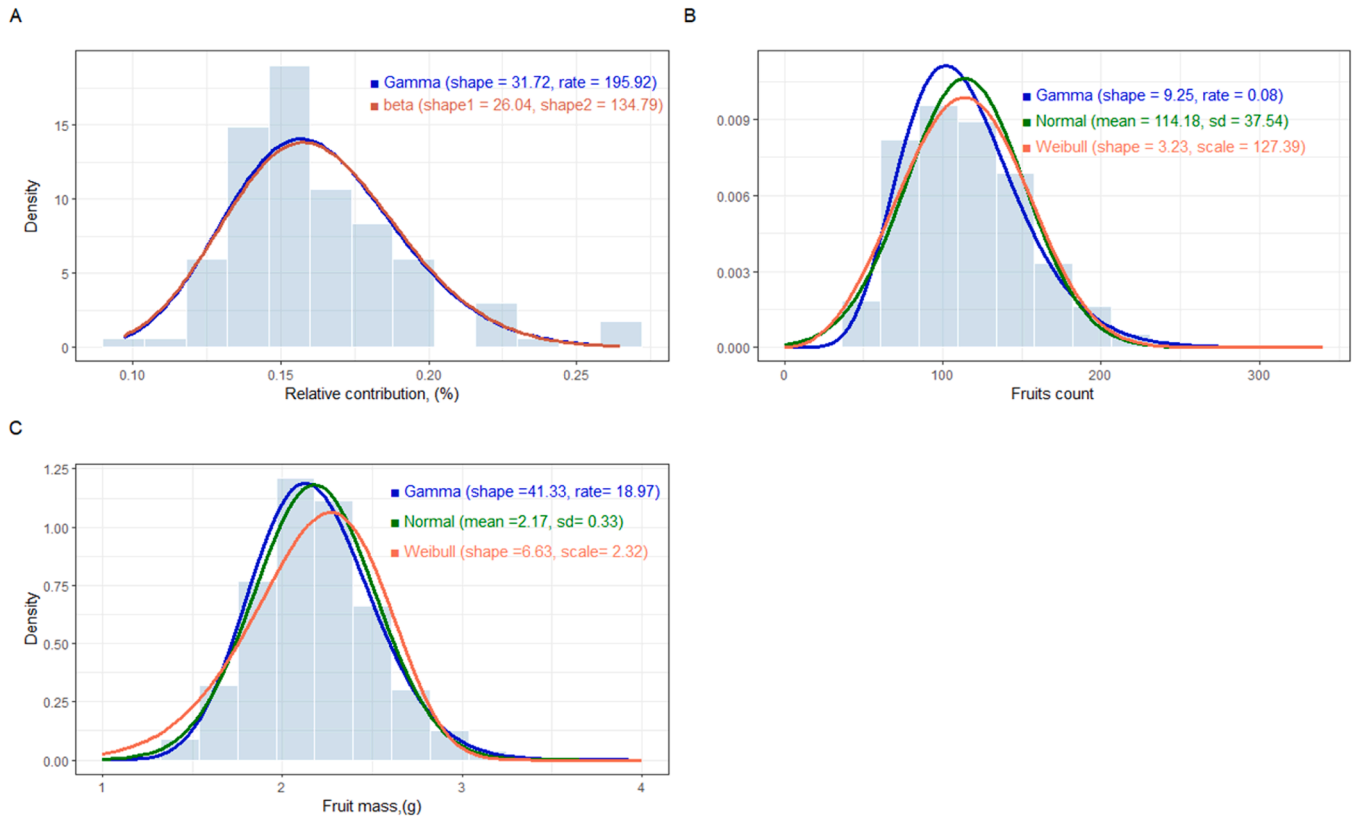

Fig. 3. Density and fit of the probabilistic distributions that indicate the behavior of the contribution of the productive branches evaluated in the tree ($CPBT_i$) (A), the number of fruits per branch predicted by Machine Learning (FB_{ml}) (B) and the individual fruit mass (FW) (C) for the first and second years of crop production.

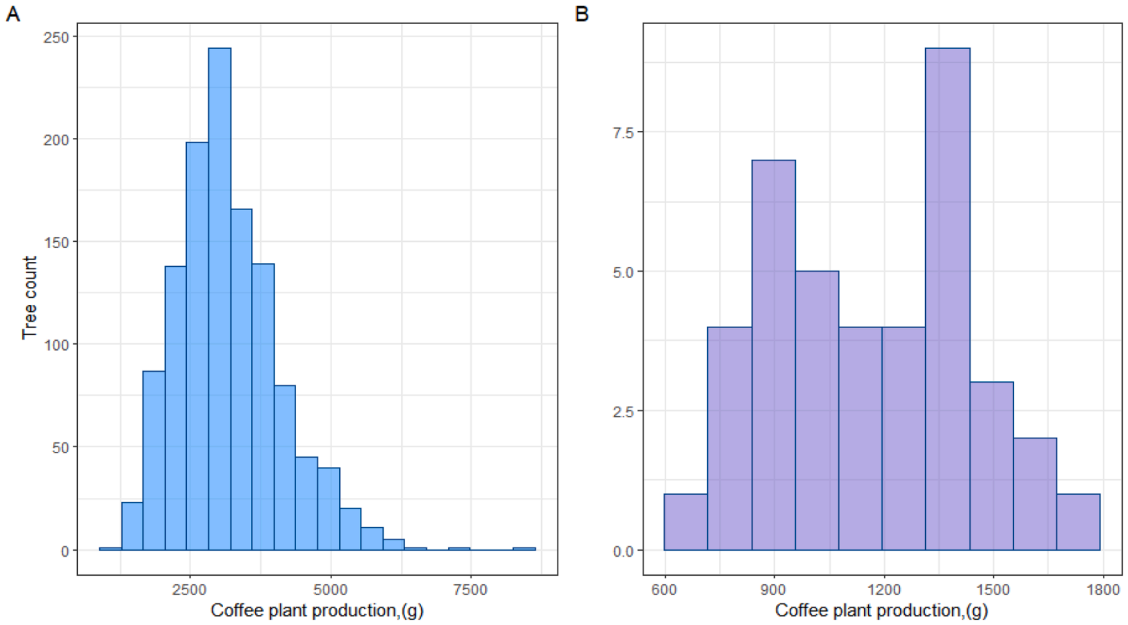

Fig. 4. Histograms presenting the distribution per tree in the simulated (blue bars) (A) and observed productions (purple bars) (B) for the first production year.

Unlike RF and SVM models, which are among the most commonly used techniques in crop yield prediction [42], ANN allows a more flexible representation of interactions between variables, integrating combined effects and dependencies that are common in crop ecophysiological processes [41]. Thus, the ANN demonstrated not only a greater predictive capacity than traditional methods but also provided a relevant contribution to the interpretation of the general structure of the productive capacity of coffee plants.

The standardization of the simulation for the yields based on the parameters FB_{ml} obtained from the ANN model and the FM measure is modeled with greater precision through a gamma distribution than through the normal and Weibull distributions (Fig. 3b and Fig. 3c). These results are consistent with the theoretical properties of the gamma distribution [43], which is defined by its flexibility to represent

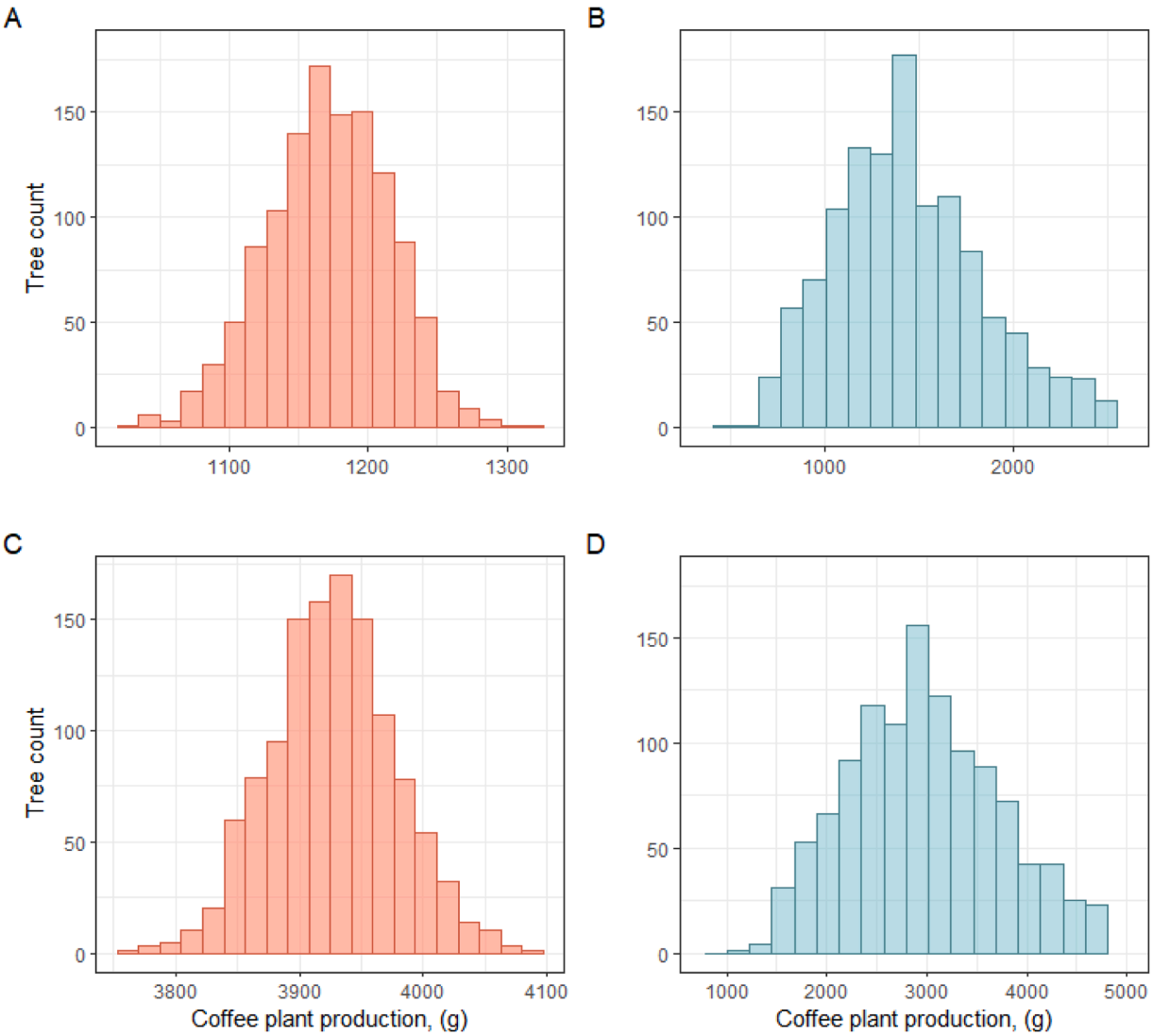

Fig. 5. Distributions of coffee cherries production observed per tree through the bootstrap method (red bars) and those obtained by simulation (blue bars). AB represents the data from the first production year, and CD represents the data from the second production year.

positively skewed continuous distributions, its restriction to positive values, and its ability to simultaneously capture the central tendency of the density and the behavior of the right tail, where extreme empirical values are concentrated. Although the estimated averages were similar between the types of probabilistic distributions, the normal distribution, owing to its assumption of symmetry and thin tails, failed to reproduce the empirical bias or the real dispersion of the observations, whereas the Weibull distribution, despite allowing different bias configurations, presented limitations in accurately capturing the global curvature of the density function and the rate of decay in the tail region, which modified the statistical criteria for the selection of the types of distributions, Table 1.

In the case of CPBTi, the beta distribution was determined in this study to be the most appropriate because its domain was bounded between 0 and 1 (Fig. 3a), which was consistent with the percentage nature of the variable. In addition, the beta distribution has been reported as highly versatile for modeling different degrees of asymmetry, the concentration of mass in specific intervals, and variability both in the center and at the extremes of the distribution [44]. These results reveal that both the beta and gamma distributions were robust in terms of the description of the probabilistic behavior of the analyzed variables, allowing a more realistic representation of the empirical form of the observed densities, as corroborated in biological processes of the production of various crops [12,13].

Prior to the simulation of crop yields, in this study, it was demonstrated that the production per plant (potential yield) can be operationally simulated from the number of fruits in the two most productive branches of the tree. Simulation of production with respect to the total number of productive branches is performed, the estimate of the yield per tree tends to be above the observed empirical average (Fig. 4). The observed discrepancies suggest that, although year-specific correction factors were used, they did not fully capture the interannual changes in the structure–yield relationship between the first and second production years. This may be due to year-to-year variation in fruit distribution and branch productivity efficiency, which would explain the overestimation in the first year and the underestimation in the second year [18,20]. This overestimation corresponds to the intraspecific heterogeneity of the crop, since not all branches have the same productive capacity; for example, variations in the distribution of the number of fruits are due to morphological and physiological changes caused by interactions with climatic conditions and crop management [22,45]. Nevertheless, this methodological approach may present limitations when extrapolated to production systems with different planting densities, pruning strategies, or environmental conditions, where the distribution of fruit production among branches may differ substantially.

The selection of the two most productive branches should be

interpreted as an operational simplification that proved effective under the conditions evaluated in this study, but which may require adjustment for broader application contexts. Therefore, this approach represents a robust starting point for the development of estimation models of yield per tree, which can be refined by incorporating agronomic and climatic covariables and statistical or Bayesian modeling techniques to capture intraspecific variability and thus reduce estimation error [4,9]. More broadly, these findings highlight the inherent difficulty of representing complex biological production systems through simplified structural assumptions, a challenge that is common across crop species and production environments.

To improve the agreement between simulated and empirical yields, a correction factor was incorporated into the production estimates (Fig. S3). This factor was not implemented as a post-hoc calibration step. Instead, it was defined as a predefined adjustment to account for systematic deviations associated with the empirical variability inherent to field conditions and tree-level production data (Fig. 5). By integrating this adjustment into the simulation framework, the model achieved a more consistent representation of the observed production per tree. The approach reduced bias while preserving the underlying distribution patterns. Consequently, the correction factor functioned as a statistical harmonization mechanism that improved the external validity of the simulations by aligning both the central tendency and the dispersion of simulated yields with the observed values [46].

Similarly, the implementation of *bootstrap* resampling techniques based on empirical information is key where the availability of data is limited by logistical, operational or resource constraints, allowing us to approximate the sample distribution of production without resorting to strict parametric assumptions and strengthening the quantification of the uncertainty associated with the estimates [38]. Together, the adjustment by the correction factor and the use of the *bootstrap* provide a robust framework to evaluate the degree of approximation between the distribution of simulated production and that observed (Fig. 5), constituting a reliable starting point for the validation of the algorithm.

5. Conclusions

Integration of machine learning techniques with a probabilistic approach suggests potential for predicting fruit load and simulating coffee crop yields from non-destructive vegetative measurements. The developed framework successfully represented production variability within the empirical range observed at the tree level, supporting its applicability for crop yield estimation under field conditions. Among the evaluated models, the ANN approach showed the best performance for estimating fruit load and correcting residual estimation error from aerial measurements of vegetative growth in coffee Castillo® Centro variety plants. Future studies should address the incorporation of complementary information at the tree and plot levels to characterize spatial and productive variability in greater detail, which could contribute to improved calibration processes and a more accurate representation of field conditions.

Funding sources

This work was supported by National Coffee Research Center (Cenicafé) (Crossref Funder ID 100,019,597), project FIT102032, funded by Colombian Federation of Coffee Growers-FNC.

Statement for studies in humans/animals

This article does not contain any studies with human participants or animals performed by any of the authors.

CRediT authorship contribution statement

Luis Carlos Imbachí-Quinchua: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Carlos Andrés Unigarro:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Álvaro Gaitán-Bustamente:** Writing – review & editing, Funding acquisition, Formal analysis, Conceptualization. **Andrés Felipe León-Burgos:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Colombian Federation of Coffee Growers (FNC in Spanish) and National Coffee Research Center (Cenicafé in Spanish) for funding this study. Thanks, Alexander Jaramillo Jiménez, José Raúl Rendón and Jhon Félix Trejos Pinzon for your support for measurements in field and to Ninibeth Sarmiento for help with the meteorological data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2026.102256.

Data availability

Data will be made available on request.

References

[1] Y. Patil, H. Ramachandran, S. Sundararajan, P. Srideviponmalar, Comparative analysis of machine learning models for crop yield prediction across multiple crop types, SN. Comput. Sci. 6 (1) (2025) 1–15, https://doi.org/10.1007/s42979-024-03602-w.

[2] C. Trentin, Y. Ampatzidis, C. Lacerda, L. Shiratsuchi, Tree crop yield estimation and prediction using remote sensing and machine learning: a systematic review, Smart Agric. Technol. 9 (2024) 100556, https://doi.org/10.1016/j.atech.2024.100556.

[3] S. Castro-Tanzi, M. Flores, N. Wanner, T.V. Dietsch, J. Banks, N. Ureña-Retana, M. Chandler, Evaluation of a non-destructive sampling method and a statistical model for predicting fruit load on individual coffee (*Coffea arabica*) trees, Sci. Hortic. 167 (2014) 117–126, https://doi.org/10.1016/j.scienta.2013.12.013.

[4] C.H. Freitas, R.D. Coelho, J. de Oliveira Costa, P.C. Sentelhas, Equationing Arabica coffee: adaptation, calibration, and application of an agrometeorological model for yield estimation, Agric. Syst. 223 (2025) 104181, https://doi.org/10.1016/j.agsy.2024.104181.

[5] S.O. Araújo, R.S. Peres, J.C. Ramalho, F. Lidon, J. Barata, Machine learning applications in agriculture: current trends, challenges, and future perspectives, Agronomy 13 (12) (2023) 2976, https://doi.org/10.3390/agronomy13122976. Article.

[6] B.M. Lionel, R. Musabe, O. Gatera, C. Twizere, A comparative study of machine learning models in predicting crop yield, Discov. Agric. 3 (1) (2025) 1–30, https://doi.org/10.1007/s44279-025-00335-z.

[7] A. Júnior, C.A.M. de, G.D. Martins, L.C.M. Xavier, B.S. Vieira, R.B.A. Gallis, E. F. Fraga Junior, R.S. Martins, A.P.B. Paes, R.C.P. Mendonça, J.V.N. Lima, Estimating coffee plant yield based on multispectral images and machine learning models, Agronomy 12 (12) (2022) 3195, https://doi.org/10.3390/agronomy12123195.

[8] M. Martello, J.P. Molin, M.C.F. Wei, R. Canal Filho, J.V.M. Nicoletti, Coffee-yield estimation using high-resolution time-series satellite images and machine learning, AgriEngineering 4 (4) (2022) 888–902, https://doi.org/10.3390/agriengineering4040057.

[9] C.H. Freitas, R.D. Coelho, J.O. Costa, P.C. Sentelhas, Smart coffee: machine learning techniques for estimating Arabica coffee yield, AgriEngineering 6 (4) (2024) 4925–4942, https://doi.org/10.3390/agriengineering6040281.

[10] K.G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine Learning in agriculture: a review, Sensors 18 (8) (2018) 2674, https://doi.org/10.3390/s18082674.

[11] R. Montes-Pajuelo, Á.M. Rodríguez-Pérez, R. López, C.A. Rodríguez, Analysis of probability distributions for modelling extreme rainfall events and detecting

climate change: insights from mathematical and statistical methods, Mathematics 12 (7) (2024) 1093, https://doi.org/10.3390/math12071093.
[12] W.D. Reynolds, C.F. Drury, L.A. Phillips, X. Yang, I.V. Agomoh, An adapted Weibull function for agricultural applications, Can. J. Soil Sci. 101 (4) (2021) 680–702, https://doi.org/10.1139/cjss-2021-0046.
[13] M. Tesfa, T. Zewotir, S.A. Derese, D.B. Belay, M. Laing, Genotype selection for grain yield of sorghum through generalized Linear Mixed model, Agronomy 13 (3) (2023) 852, https://doi.org/10.3390/agronomy13030852.
[14] H.-J. Bak, E.-J. Kim, J.-H. Lee, S. Chang, D. Kwon, W.-J. Im, D.-H. Kim, I.-H. Lee, M.-J. Lee, W.-H. Hwang, N.-J. Chung, W.-G. Sang, Canopy-level rice yield and yield component estimation using NIR-based vegetation indices, Agriculture 15 (6) (2025) 594, https://doi.org/10.3390/agriculture15060594.
[15] A. Jaramillo-Robledo, El clima de la caficultura en Colombia. Colección Libros 80 Años Cenicafé, 2018, https://doi.org/10.38141/cenbook-0031.
[16] F.M. DaMatta, S.C.V. Martins, J.D.C. Ramalho, Ecophysiology of coffee growth and production in a context of climate changes. En *Advances in Botanical Research*, Academic Press, 2025, https://doi.org/10.1016/bs.abr.2024.07.004.
[17] C.A. Unigarro, D.G. Cayón Salinas, A.F. León-Burgos, C.P. Flórez-Ramos, Flowering and fruiting of coffea arabica L.: a comprehensive perspective from phenology, Plants 14 (21) (2025) 3396, https://doi.org/10.3390/plants14213396.
[18] A.F. León-Burgos, J.R.R. Sáenz, L.C.I. Quinchua, M.A. Toro-Herrera, C.A. Unigarro, V. Osorio, H.E. Balaguera-López, Increased fruit load influences vegetative growth, dry mass partitioning, and bean quality attributes in full-sun coffee cultivation, Front. Sustain. Food Syst. 8 (2024) 1379207, https://doi.org/10.3389/fsufs.2024.1379207.
[19] A.F. León-Burgos, J.R.R. Sáenz, L.C.I. Quinchua, C.A. Unigarro, V. Osorio, S. S. Khalajabadi, H.E. Balaguera-López, Varying fruit loads modified leaf nutritional status, photosynthetic performance, and bean biochemical composition of coffee trees, Sci. Hortic. 329 (2024) 113005, https://doi.org/10.1016/j.scienta.2024.113005.
[20] C.A. Unigarro, L.M.D. Bejarano, J.R. Acuña, Effect of fruit load of the first coffee harvests on leaf gas exchange, Pesquisa Agropecuária Trop. 51 (2022) e69865, https://doi.org/10.1590/1983-40632021v5169865.
[21] W.L. Almeida, R.T. Ávila, J.P. Pérez-Molina, M.L. Barbosa, D.M.S. Marçal, R.P.B. de Souza, P.B. Martino, A.A. Cardoso, S.C.V. Martins, F.M. DaMatta, The interplay between irrigation and fruiting on branch growth and mortality, gas exchange and water relations of coffee trees, Tree Physiol. 41 (1) (2021) 35–49, https://doi.org/10.1093/treephys/tpaa116.
[22] M. Rakocevic, M.B. dos Santos Scholz, R.A.A. Pazianotto, F.T. Matsunaga, J. C. Ramalho, Variation in yield, berry distribution and chemical attributes of coffea arabica beans among the canopy strata of four genotypes cultivated under contrasted water regimes, Horticulturae 9 (2) (2023) 215, https://doi.org/10.3390/horticulturae9020215.
[23] A.D. Bote, V. Jan, Branch growth dynamics, photosynthesis, yield and bean size distribution in response to fruit load manipulation in coffee trees, Trees 30 (4) (2016) 1275–1285, https://doi.org/10.1007/s00468-016-1365-x.
[24] F.M. DaMatta, R.L. Cunha, W.C. Antunes, S.C.V. Martins, W.L. Araujo, A.R. Fernie, G.A.B.K. Moraes, In field-grown coffee trees source–sink manipulation alters photosynthetic rates, independently of carbon metabolism, via alterations in stomatal function, New Phytol. 178 (2) (2008) 348–357, https://doi.org/10.1111/j.1469-8137.2008.02367.x.
[25] P. Vaast, J. Dauzat, M. Génard, Modeling the effects of fruit load, shade and plant water status on coffee berry growth and carbon partitioning at the branch level, Acta Hortic. 584 (2002) 57–62, https://doi.org/10.17660/ActaHortic.2002.584.5.
[26] J.C. Chrispell, E.W. Jenkins, K.R. Kavanagh, M.D. Parno, Characterizing prediction uncertainty in agricultural modeling via a coupled statistical–Physical framework, Modelling 2 (4) (2021) 753–775, https://doi.org/10.3390/modelling2040040.
[27] D. Batool, M. Shahbaz, H. Shahzad Asif, K. Shaukat, T.M. Alam, I.A. Hameed, Z. Ramzan, A. Waheed, H. Aljuaid, S. Luo, A hybrid approach to tea crop yield prediction using simulation models and machine learning, Plants 11 (15) (2022) 1925, https://doi.org/10.3390/plants11151925.
[28] J.R. Rendón, Administración de sistemas de producción de café a libre exposición solar. En *Manejo Agronómico De Los Sistemas de Producción de Café*, Centro Nacional de Investigaciones del Café, 2020, pp. 34–71, https://doi.org/10.38141/10791/0002_1.
[29] Agroclima-Plataforma Agroclimática Cafetera, Website, available at. https://agroclima.cenicafe.org/, 2025. Accessed on October 2025.
[30] C.P. Flórez, J.C. Arias, H. Cortina, M. Moncada-Botero, J. Quiroga-Cardona, D. M. Molina, J.C. García-López, Variedades Castillo ® Zonales. Resistencia a la roya con mayor productividad, Avances Técnicos Cenicafé 489 (2018) 1–8, https://doi.org/10.38141/10779/0489.
[31] Centro Nacional de Investigaciones de Café. (2021). Guía más agronomía, más productividad, más calidad (3a ed.). Cenicafé. https://doi.org/10.38141/cenbook-0014.
[32] J.R. Rendón, E.C. Montoya, ¿Cómo registrar las floraciones en los cafetales? Avances Técnicos Cenicafé 455 (2025) 1–4, https://doi.org/10.38141/10779/0455.
[33] J. Arcila-Pulgarín, L. Buhr, H. Bleiholder, H. Hack, U. Meier, H. Wicke, Application of the extended BBCH scale for the description of the growth stages of coffee (Coffea spp.), Ann. Appl. Biol. 141 (1) (2002) 19–27, https://doi.org/10.1111/j.1744-7348.2002.tb00191.x.
[34] C.A. Unigarro-Muñoz, J.D. Hernández-Arredondo, E.C. Montoya-Restrepo, R. D. Medina-Rivera, L.N. Ibarra-Ruales, C.Y. Carmona-González, et al., Estimation of leaf area in coffee leaves (Coffea arabica L.) of the Castillo® variety, Bragantia 74 (2015) 412–416, https://doi.org/10.1590/1678-4499.0026.
[35] J.B. Ruhland, I. Masoudian, D. Heider, Enhancing deep neural network training through learnable adaptive normalization, Knowl. Based. Syst. 326 (2025) 113968, https://doi.org/10.1016/j.knosys.2025.113968.
[36] T. Hastie, R. Tibshirani, J. Friedman, Model assessment and selection, in: E. T. Hastie, R. Tibshirani, J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 1995, pp. 219–259, https://doi.org/10.1007/978-0-387-84858-7_7.
[37] A.C. Cullen, H.C. Frey, Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs, Springer Science & Business Media, 1999.
[38] C.P. Robert, G. Casella, Monte Carlo Integration, in: C.R. En, G. Casella (Eds.), Introducing Monte Carlo Methods with R, Springer, 2010, pp. 61–88, https://doi.org/10.1007/978-1-4419-1576-4_3.
[39] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, 2013, https://doi.org/10.1017/CBO9780511802843.
[40] Development Core Team R, R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019.
[41] Y. Kittichotsatsawat, N. Tippayawong, K.Y. Tippayawong, Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques, Sci. Rep. 12 (1) (2022) 14488, https://doi.org/10.1038/s41598-022-18635-5.
[42] M.D.A. Jabed, M.A. Azmi Murad, Crop yield prediction in agriculture: a comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability, Heliyon 10 (24) (2024) e40836, https://doi.org/10.1016/j.heliyon.2024.e40836.
[43] J. Reyes, C. Marchant, K.I. Santoro, Y.A. Iriarte, A versatile distribution based on the incomplete gamma function: characterization and applications, Mathematics 13 (11) (2025) 1749, https://doi.org/10.3390/math13111749.
[44] J.D. Haskett, Y.A. Pachepsky, B. Acock, Use of the beta distribution for parameterizing variability of soil properties at the regional level for crop yield estimation, Agric. Syst. 48 (1) (1995) 73–86, https://doi.org/10.1016/0308-521X(95)93646-U.
[45] V.O. Sadras, R.F. Denison, Do plant parts compete for resources? An evolutionary viewpoint, New Phytol. 183 (3) (2009) 565–574, https://doi.org/10.1111/j.1469-8137.2009.02848.x.
[46] S.S. Jagtap, J.W. Jones, Adaptation and evaluation of the CROPGRO-soybean model to predict regional yield and production, Agric. Ecosyst. Environ. 93 (1) (2002) 73–85, https://doi.org/10.1016/S0167-8809(01)00358-9.