Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Socioeconomic data alone can reliably predict smallholder maize yields via machine learning benchmarking in Cameroon

Sopkoutie Kengni Nerlus Gautier

gautier.sopkoutie@univ-dschang.org

University of Dschang

Bidias Aboh Francis

University of Dschang

Deffo Tchouan Gilchrist

University of Mbouoh

Kamseu Mogo Jean Paul

Agricultural Research Institute for Development

Research Article

Keywords: Maize yield, Machine Learning, Prediction, Socio-economic data, Cameroon

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11140268/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Socioeconomic data alone can reliably predict smallholder maize yields via machine learning benchmarking in Cameroon

Sopkoutie Kengni Nerlus Gautier[1], Bidias Aboh Francis[1,], Deffo Tchouan Gilchrist[2], Kamseu Mogo Jean Paul[3]

[1]University of Dschang, Faculty of Agronomy and Agricultural Sciences, Plant pathology and agricultural zoology research unit (UR-PHYZA), Dschang, Cameroon,

[2]University of Mbouo, Research and Application Unit in Plant Production (URAPP), Bandjoun, Cameroon,

[3]Agricultural Research Institute for Development, Dschang, Cameroon

Corresponding authors. gautier.sopkoutie@univ-dschang.org

Abstract

Background : Predicting maize yields is a major challenge for food security and the optimization of agricultural systems in Sub-Saharan Africa. While climate and spectral data are often prioritized, the predictive power of socio-economic factors remains less explored. This study aimed to evaluate the performance of diverse Machine Learning (ML) in predicting maize yield using farmer survey data. **Methodology :** We assessed the performance of seven ML models and two ensemble methods (Stacking, AdaBoost)(multiple linear regression (MLR), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Random Forest (RF), Artificial Neural Network (ANN), and XGBoost, alongside Stacking and AdaBoost models) using survey data collected from 354 maize farmers in Vina, Cameroon. Variable importance analysis, model training, and robust cross-validation (*k=5*) were conducted. **Results:** Variable importance analysis revealed that profit per hectare was the dominant predictor (score=0.895; r=0.85 with yield), followed by fertilizer use and agricultural expenditure. Modeling results demonstrated the superiority of ensemble methods. XGBoost achieved the best performance (R^2=0.98; RMSE=0.128), closely followed by Random Forest and Stacking, while linear models and K-Nearest Neighbors (KNN) exhibited insufficient accuracy. Cross-validation confirmed the robustness and generalization ability of ensemble models, with XGBoost, Random Forest, and Stacking all maintaining high accuracy ($R^2 \approx 0.93$-0.94) and low variance. **Conclusion:** Our findings demonstrate that socio-economic data, when analyzed by robust ensemble algorithms, can provide reliable and accurate maize yield predictions. This approach offers a cost-effective and robust alternative to traditional climate- or spectral-based modeling, presenting new opportunities for agricultural planning and precision agriculture in resource-limited areas of Sub-Saharan Africa.

Keywords: Maize yield, Machine Learning, Prediction, Socio-economic data, Cameroon.

1. Introduction

Agriculture and climate change are closely linked, as climate is one of the main factors destabilising agricultural systems (Pathak et al., 2012). Climate change contributes to reduced food availability, poorer nutritional quality and reduced access to food (Rosenberg, 1992). Variations in temperature, precipitation distribution and the occurrence of extreme events such as heat waves, diseases and pest infestations exacerbate these disruptions and affect the nutritional composition of certain food products (Subedi et al., 2023). In this context, it is crucial to evaluate and model the effects of climate in order to increase the resilience of agricultural systems in an environment characterised by pronounced variability (Ruane et al., 2013).

According to Liwumi and Ramankutty (2015), three parameters yield, cultivated area, and crop frequency form the basis of the agricultural production equation. It is estimated that approximately 70% of climate-related variations in production are due to fluctuations in cultivated area and/or crop frequency (Habib-ur-Rahman et al., 2022). Yield forecasting thus becomes essential not only for crop management, but also for trade, economic planning, food security and agricultural policy-making. While farmers once relied on experience and historical data, the advent of simulation models, big data, machine learning (ML) and high-performance computing has significantly improved the accuracy of predictions (Drummond et al., 2003; Vincenzi et al., 2011; González Sánchez et al., 2014; Jeong et al., 2016; Pantazi et al., 2016; Cai et al., 2017; Chlingaryan et al., 2018; Crane-Droesch, 2018; Basso & Liu, 2019; Shahhosseini et al., 2019).

Numerous studies have examined the effects of climate change on agricultural yields, both locally and globally, using deterministic or artificial intelligence-based approaches (Chalinor et al., 2014; Thornton et al., 2009; Alenxendrov and Hoogenboom, 2000). For example, Romeijn et al. (2016) compared deterministic methods and hierarchical analytical processes to assess agricultural land suitability, while Aschonitis et al. (2012) analysed soil vulnerability to water and nitrogen losses using regression models. Other studies also rely on deterministic or probabilistic approaches (Meenken et al., 2020; Kingsley et al., 2021). However, these methods have limitations: they are time-consuming to implement, complex and require significant resources (Sharma et al., 2020a; Sharma et al., 2020b). These constraints explain the rise of machine learning techniques, which are more flexible and can be automated. For example, Kouadio et al. (2018) used soil fertility indicators (organic matter, potassium, boron, phosphorus, nitrogen, pH, etc.) to predict Robusta coffee yields in Vietnam.

Supervised learning, widely used in agriculture, falls into two main categories: regression and classification (James et al., 2013). Regression is preferred for yield forecasting because the variable to be predicted is continuous. Its applications cover yield forecasting (Drummond et al., 2003; Vincenzi et al., 2011; González Sánchez et al., 2014; Jeong et al., 2016; Pantazi et al., 2016; Cai et al., 2017; Chlingaryan et al., 2018; Basso & Liu, 2019; Shahhosseini et al., 2019; Emirhüseyinoğlu & Ryan, 2020; Khaki et al., 2020), crop quality assessment (Hoogenboom et al., 2004; Karimi et al., 2008; Mutanga et al., 2012; Shekoofa et al., 2014), water management (Mohammadi et al., 2015; Mehdizadeh et al., 2017) and soil management (Johann et al., 2016; Morellos et al., 2016; Nahvi et al., 2016).

Recent research also shows that a single machine learning model can be outperformed by a set of models, known as ensemble learning (Zhang & Ma, 2012). This strategy is effective because it reduces bias, variance, or both, and better represents the actual distribution of data, provided that the base models are sufficiently diverse (Dietterich, 2000; Pham & Olafsson, 2019a, 2019b; Shahhosseini et al., 2019a, 2019b). The use of ensemble learning to address agricultural and ecological issues is thus growing significantly.

This approach is becoming increasingly widespread: bagging and, in particular, random forests (Vincenzi et al., 2011; Mutanga et al., 2012; Fukuda et al., 2013; Jeong et al., 2016), boosting (Sajedi-Hosseini et al., 2018) and stacking (Cai et al., 2017; Shahhosseini et al., 2019) are a few examples. However, to our knowledge, few studies directly compare the effectiveness of different ensemble methods in the agricultural field, particularly when the data exhibit spatio-temporal correlations.

Various machine learning algorithms have been used to predict agricultural yields: random forest (RF) (Han et al., 2020), partial least squares (PLS) (Maimaitijiang et al., 2017), ridge regression (RR) (Ahmed et al., 2022), k-nearest neighbours (KNN) (Feng et al., 2020), and extreme gradient boosting (XGBoost) (Babaie Sarijaloo et al., 2021). However, the performance of these models varies greatly depending on crops and environments, due to the quality and representativeness of the data, as well as the interdependencies between variables (Maseko et al., 2024). In the presence of bias or overfitting, the accuracy of predictions deteriorates significantly (Montesinos López et al., 2022).

In recent years, the rise of big data has led to the widespread use of machine learning methods in agriculture (Zhang et al., 2021), which often outperform traditional linear regression approaches. Deep learning (DL) techniques, based on the stacking of non-linear layers (LSTM, DNN, CNN, RNN), have demonstrated significantly higher accuracy (Cai et al., 2019). To overcome certain persistent limitations, ensemble learning is a promising alternative, as it integrates data fusion, modelling and exploration into a unified framework (Tao et al., 2025).

Among its variations, stacked regression combines several predictive models to improve accuracy (Zhang et al., 2015; Zhang et al., 2022), while feature-weighted ensemble assigns weight to variables based on their correlation with the target (Liu et al., 2020; Niño-Adan et al., 2021).

In this study, we propose a hybrid method based on feature-weighted ensemble learning to modulate the predictions of base models according to their individual performance, before training a meta-learner. To further optimise accuracy, a third layer is added: a simple mean ensemble, which aggregates the predictions from stacking and weighting, then compares them to the observed values.

The data used comes from maize producers in the Vina division of the Adamaoua region (Cameroon). The objective of this work is to evaluate the performance of different machine learning models in predicting maize yields and to bring new technical perspectives to precision agriculture in Cameroon.

2. Methodology

2.1. Study area and data collection

The study was conducted on a sample of 354 corn producers in the district of the Vina, Cameroon. Information collected through a questionnaire (see Supplementary File 1), previously developed for this study, was grouped into 30 explanatory variables, organized into two main categories: socio-economic characteristics and farming practices. These variables include data on the farmer's profile (age, gender, education level, experience) and farming methods (cropping system, acreage, type of fertilization, weed control methods). Yield (expressed in kilograms per hectare) was selected as the dependent variable. This structure provided a rich and heterogeneous database, suitable for the application of machine learning algorithms. The data was processed and prepared using *open-source* Python tools. Machine learning algorithms were implemented for modeling, and robust techniques such as five-fold cross-validation were used to evaluate the models and the importance of the variables based on the *feature importance* script. Key steps included data processing, data preparation and variable importance assessment, model selection, and performance evaluation using metrics tailored to the study objectives.

2.2. Data preprocessing

Data preparation began with the compilation and structuring of information from 354 corn producers in an Excel 2019 file (Ntsoli et al., 2024). This file was then imported into Python using the *pandas* library and converted into a *DataFrame* object, providing a matrix organization suitable for statistical processing. An initial check of the completeness and consistency of the data was carried out following

the approach described by Hair et al. (2022), which allowed missing values to be detected using the *isnull()* function, duplicates to be identified using *duplicated*(), and deleted using *drop_duplicates*(). Missing values were handled in accordance with the recommendations of Tabachnick & Fidell (2019). Individuals with too many missing values were excluded from the dataset, while other missing values were imputed using the mean or median for continuous variables and the mode for categorical variables, using *pandas* fill functions. Outliers were then identified using graphical and statistical approaches. Visually, boxplots and histograms were generated with *matplotlib,* while *z-scores* were calculated using the *zscore* function from the *scipy* library, in accordance with the approach proposed by Jolliffe & Cadima (2016). The detected outliers were treated by winsorization, adjusting the data to the 1st and 99th percentile thresholds, which reduced their influence without removing them. After cleaning, the data were transformed and coded to facilitate their use in statistical models. Nominal categorical variables were converted to binary indicators using *pandas' get_dummies*(), while ordinal variables were encoded with *scikit-learn's LabelEncoder* class, as recommended by Kassambara (2017). Continuous variables were standardized () using *StandardScaler*, and normalizations were performed using *MinMaxScaler* available in the *sklearn.preprocessing* module. The latter was preferred in the case of distributions influenced by extreme values, as it is based on the median and interquartile range (Pedregosa et al., 2011). To correct certain asymmetries, logarithmic and square root transformations were also applied using the *numpy* functions log1p () and sqrt (), in accordance with Field's (2018) recommendations. An exploratory analysis was then conducted to highlight the relationships between variables. The distributions were represented as histograms and boxplots (Figure 2) using *matplotlib,* and a correlation matrix was calculated using *pandas.*

2.3. Variable selection

This was performed after training the Random Forest model, using the intrinsic *feature_importances_* property, which calculates importance based on node impurity. this approach quantifies the importance of each variable by measuring the average decrease in impurity (*gini or entropy*) obtained when a feature is used to divide nodes across all trees in the forest. The importances were extracted from the trained model, normalized to sum to 1, and then sorted in descending order to identify the most influential variables for predicting the target variable. For a robust interpretation, the results were visualized as a vertical bar chart (Figure 2), allowing clear identification of the dominant characteristics and validation of their relevance consistent with the field of study, thus confirming the logic of the model and guiding possible future variable selections. The Pearson correlation coefficient was used to analyze the relationship between various influencing factors and rapeseed yield, and the significance test was carried out (Figure 3). On this basis, variables with a significant correlation ($p < 0.05$) were selected for comprehensive analysis.

2.4. Definition of machine learning (ML) models

To predict corn yield (RDT) from data collected from 354 farmers, in accordance with study of Yang et al. (2024) , nine (09) machine learning algorithms were used in this work: linear regression, Random Forest, SVR, RF, Ridge, KNN, and XGBoost. The aim was to assess their ability to handle complex relationships between socio-economic and agronomic variables. The methodological choice adopted in this study responds to the need to reconcile scientific rigor and applicability in the African agricultural context. The use of socio-economic data, often neglected in favor of climatic or spectral data, is justified by its accessibility and operational relevance in resource-limited areas (Fan et al., 2023 ; Gao et al., 2023). To ensure a comprehensive assessment, a diverse set of models ranging from linear regression to advanced ensemble methods (XGBoost, Stacking) was used, in line with recent recommendations highlighting their superiority in terms of robustness and generalization (Chen et al., 2025 ; Li et al., 2023). Finally, the systematic application of data preparation techniques and five-fold cross-validation reinforce the reliability of the results and minimize the risks of bias or overfitting (Hair et al., 2022 ; Yang et al., 2024).

2.4.1. Individual models

Linear regression

The linear regression is one of the simplest and most common machine learning algorithms (Maulud & Abdulazeez, 2020). Linear regression is a linear method of constructing models that describes the relationship between a single output variable (dependent variable) and multiple input explanatory variables (independent variable). The types mainly included are univariate linear regression predicts a quantitative response based on multiple predictor variables. A multiple linear regression model is in the following form (James et al., 2013).

$$Y = b0 + b1X1 + b2X2 + \ldots + bpXp + \epsilon$$

in which Y is the response variable, Xj are the independent variables, bj are the coefficients, and ϵ is the error term. The coefficients are estimated by minimizing the loss function L, as shown below. where yi ˆis the prediction for yi.

$$y=b1x1+\ldots+bnxn+b0$$

Where, $x1 \ldots xi$ are the natural and socio-economic indicators, y is the rapeseed yield, $b1 \ldots bn$ is the regression coefficient, and b0 is a constant term

Kernel Ridge Regression (KRR)

193 Ridge regression (RR) is a simple yet effective non-linear regression method for prediction, particularly
194 when combined with a kernel, resulting in kernel ridge regression (KRR). The latter allows the input
195 data, derived from a nonlinear time series, to be projected from a low-dimensional space to a high-
196 dimensional space (Naik et al., 2018; Li et al., 2019). The kernel function acts as a feature mapping
197 defined in a Hilbert space Hk of dimension d, $\Psi : X \rightarrow Hk$, telle que $k\,(xi, xj) = \{\Psi\,(xi), \Psi\,(xj)\}\ Hk$. In this
198 study, we follow Li et al. (2019) to implement KRR. With kernel functions and n data
199 samples$(x1, y1), (x2, y2), \ldots, (xn, yn) \in X * Y$ (yi is the target value corresponding to xi, $i = 1, 2, \ldots, n$),
200 the kernel matrix equation is:

201 $$K=\begin{bmatrix} k(x1,x1) & k(x1,x2)\ldots k(x1,xn) \\ k(x1,x2) & k(x1,x2)\ldots k(x2,xn) \\ k(xn,x1) & k(xn,x2)\ldots k(xn,xn) \end{bmatrix}$$

202 The KRR problem can be formulated as :

203 $$\min_{w} \|Y\text{-}Kw\|^2+\lambda\|w\|^2$$

204 Here, Y is the target vector of all n data samples, w is the unknown vector, In is an identity matrix of
205 dimension n* n, and the regularization term w = λ (≥ 0 prevents w from becoming too larg

206 $$w=(K+\lambda In)^{-1}Y$$

207 **Random Forest**

208 Random Forest RF is an ensemble algorithm that combines multiple decision trees to produce robust
209 and accurate predictions (Breiman, 2001). In regression, the final prediction is the average of the trees'
210 predictions, reducing variance while maintaining low bias. Each tree is trained on a random subset of
211 the data (bootstrap) and considers a random subset of features at each node, which increases the
212 diversity of the trees and limits overfitting. A decision tree starts with a set of examples X and attempts
213 to partition them based on the values of the features xi. At each node t, the algorithm selects the feature
214 xj and threshold θ that minimize a measure of impurity or heterogeneity, such as entropy or the Gini
215 index. The Gini index is calculated as follows:

219 $$G(t)=1-\sum_{i=1}^{k} P^2 i$$

216 Where pi is the proportion of examples belonging to class i in node t and k is the number of classes.
217 Once all trees are constructed, each tree generates a prediction hk(x) for a new input x. In classification,
218 the final class y is determined by majority vote:

220 $$y=\{h1(x),h2(x),\ldots\ldots,hx(x)\}$$

221 Where T is the number of trees in the forest. In regression, predictions are averaged:

222 $$y=\frac{1}{T}\sum_{K=1}^{T} h_k(x)$$

This aggregation process is key to the robustness of RF, as it mitigates the biases of individual trees and reduces noise. The Random Forest Regressor model from scikit-learn was initialized with the following parameters: (1000 trees, default parameters) as an ensemble Baseline.

EXtreme Gradient Boosting (XGB)

XGBoost (Extreme Gradient Boosting) is a powerful machine learning technique introduced by Chen and Guestrin (2016). It is widely used for regression and classification tasks. The model is based on the principle of ensemble learning, combining several decision trees to iteratively improve the predictive accuracy of the model. By assigning weights to individual trees according to their performance, XGBoost effectively minimizes prediction errors and optimizes model generalization (Friedman, 2001). The prediction process can be defined as follows:

$$\widehat{yi}=\sum_{k=1}^{K} fk(xi),\ fk\in F$$

Where:

- $\hat{y}_i$ Represents the predicted value for observation i,
- K Is the total number of decision trees,
- fk Denotes a prediction function belonging to the set of decision trees F,
- xi Corresponds to the vector of explanatory variables for observation i.

Each new tree is added sequentially in order to correct the residual errors made by the previous trees, thereby gradually increasing the overall performance of the model.

Each tree is defined recursively as follows:

$$f_k=w_q(x;\theta_k)$$

Where w represents the leaf weights and $q(x; \theta k)$ is a function that associates the input variables $x=(x1, x2, x3, x4)$ with a leaf of tree k.θk represents the parameters of the kth tree. The final prediction is the sum of all the predictions of the trees.

$$Y=\sum_{k=1}^{K} w_q(x,\theta_k)$$

Where Y is the final prediction, K is the set of trees, andWq(x; θk) is the result of tree k.

The objective function of the XGB model generally includes a loss term, intended to minimize the difference between predicted values and actual values, as well as regularization terms to control the complexity of the trees:

$$L(\Theta)=\sum_{i=1}^{N} L(yi,\widehat{yi})+\sum_{k=1}^{K}\Omega(f_k)+\lambda\,\|\Theta\|^2$$

Where:

— N : number of training samples,

— L : loss function,

— yi : actual maize yield for sample i,

— $\widehat{yi}$: predicted yield for sample i,

— K : number of trees,

— Ω (f_k): regularization term for tree k,

— λ : regularization parameter,

— $\|\Theta k\|^2$: norm$L2$ of the parameters of tree k.

The model parameters Θk are learned by optimizing the objective function L (Θ) using techniques such as *gradient boosting*.

Support Vector Regression (SVR)

SVR is a machine learning model that extends *Support Vector Machines (SVM)* to regression tasks. Its objective is to find a regression hyperplane that minimizes the distance between the hyperplane and the data points. To handle complex relationships, SVR frequently uses kernel functions that project the data into a higher-dimensional space, where a linear regression model can be applied (Cheng et al., 2022; Kharal et al., 2024).

The fundamental SVR algorithm is defined as follows:

$$\frac{1}{2}\|W\|^2+C\sum_{i=1}^{N}(\xi_i+\xi_i^*)$$

Under the following constraints, when $y=W\cdot x+by$

$$y_i-W.xi-b\leq\xi+\xi_i$$

$$W.xi+b-yi\leq\xi+\xi i^*$$

$$\xi_i,\xi i^*\geq 0$$

Where:

— W représente les poids, represents the weights

— C, It is a regularization parameter,

— ξi* are *slack variables*,

— ε represents the tolerance margin,

— xi and yi are respectively the input variables and the output value (target).

Artificial Neural Network (ANN) Model

Artificial neural networks (ANNs) are non-linear, data-driven, and self-adaptive methods, unlike traditional model-based approaches (Zhang et al., 1998). ANNs have the ability to identify complex

relationships between input variables and corresponding target values. They can solve problems involving nonlinear and complex data, even when that data is noisy or incomplete, because they mimic the learning process of the human brain. They are therefore particularly well suited to modeling agricultural data, which is generally complex and often nonlinear.

The output of a neural network can be expressed by the following equation (Hornik et al., 1989)

$$y_t=\alpha 0+\sum_{j=1}^{n}\alpha jf(\sum_{i=1}^{m}\beta ijy_{t-1}+\beta_{oj})+\varepsilon_t$$

Where :

- y_t is the output of the neural network model (e.g., yield per plant),
- n is the number of neurons in the hidden layer,
- m is the number of neurons in the input layer,
- f is the activation function,
- $\beta ij(i=1,2,\ldots,m;j=0,1,\ldots,n)$ are the weights connecting the input neurons to those in the hidden layer ;
- $\alpha j(j=0,1,\ldots,n)$ are the weight vectors connecting the hidden layer to the output,
- $\alpha 0$ et $\beta 0j$ are the bias weights.

The activation function is a differentiable function used to smooth the result of the weighted product of the inputs and neurons. It determines the output of a neuron from one or more inputs. In this study, the logistic function was used as the activation function, while the Levenberg–Marquardt (LM) learning algorithm was used to adjust the weights in feedforward multilayer networks (Hagan & Menhaj, 1994).

2.4.2. Ensemble methods

Stacked generalization

Stacked generalization aims to minimize the generalization error of some ML models by performing at least one more level of learning task using the outputs of ML base models as inputs and the actual response values of some part of the data set (training data) as outputs (Wolpert, 1992). Stacked generalization assumes the data to be IID (independent and identically distributed) and performs a k-fold cross-validation to generate out-of-bag predictions for the validation set of each fold. Collectively, the k out-of-bag predictions create a new training set for the second level learning task, with the same size as the original training set (Cai et al., 2017). However, here the IID assumption of the data does not hold, and we cannot use k-fold cross-validation to generate out-of-bag predictions. To work around this issue, blocked sequential procedure (Cerqueira et al., 2017 ; Oliveira et al., 2019) was used to generate inputs of the stacked generalization method only using past data of some ML models by performing at least one more level of learning task using the outputs of ML base models as inputs and

the actual response values of some part of the data set (training data) as outputs (Wolpert, 1992). Stacked generalization assumes the data to be IID and performs a k-fold cross-validation to generate out-of-bag predictions for validation set of each fold. Collectively, the k out-of-bag predictions create a new training set for the second level learning task, with the same size of the original training set (Cai et al., 2017). Base Estimators: We selected a set of diverse regression models that we had previously trained as our base estimators. In our case, these included: Random Forest Regressor, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) Regressor, Linear Regression, XGBoost Regressor, AdaBoost Regressor, Kernel Ridge Regressor (Note: The ANN model was not included as a base estimator in the Stacking Regressor due to potential compatibility or training complexities within the stacking framework). Training the Base Estimators: Each of these base estimator models is trained on the original training data (specifically, the top 10 features in our case).

Generating Out-of-Fold Predictions: This is a crucial step in standard stacking to prevent the meta-regressor from overfitting to the base estimators' training data. The training data is typically split into several folds (like in cross-validation). For each fold, the base estimators are trained on the other folds and then make predictions on the held-out fold. This process is repeated for all folds, resulting in a full set of **8** "out-of-fold" predictions for the entire training dataset, generated by base estimators that did not see that specific data during their training.

Training the Meta-Regressor: The out-of-fold predictions from the base estimators become the new "features" for training the meta-regressor. The meta-regressor is trained on these out-of-fold predictions (as input) and the original target variable (as the output). In our case, we used a Linear Regression model as the meta-regressor. The meta-regressor learns how to optimally combine the predictions of the base estimators.

Making Final Predictions: When making predictions on new, unseen data (like the test set): Each of the trained base estimators makes a prediction on the new data. These predictions from the base estimators are then fed as input to the trained meta-regressor. The meta-regressor uses these base predictions to make the final stacked prediction.

2.5. Running and Configuring Machine Learning Models

The dataset was divided into two subsets according to an **80-20** ratio, meaning that **80%** of the data was used to **train the model** and **20%** was used **to evaluate** it.

The machine learning (ML) models were run using the **default settings** for each algorithm, taking into account the **computing capabilities** of the available hardware. This computing power allowed standard configurations to be maintained without imposing additional constraints on **hyperparameter**

definition**.** This approach is consistent with **recommended practices** when the main objective is to **compare several algorithms** on the same dataset (Raschka, 2018).

2.6. Machine learning model parameterization

In order to evaluate the performance of different supervised learning algorithms in modeling variables, several models were parameterized and tested as follows:

- **SVR (Support Vector Regression)** : Four kernel functions were examined: sigmoid, polynomial, linear, and RBF with an insensitivity parameter set to $\epsilon = 0.1$ and a regularization parameter ($C = 1.0$). These values were chosen for their proven effectiveness in the nonlinear modeling of agroclimatic variables (Smola & Schölkopf, 2004 ; Awad & Khanna, 2015).
- **RF (Random Forest)**: A forest of 1,000 decision trees was constructed, with each tree allowed to grow until pure leaves were obtained. This model is renowned for its robustness and ability to handle non linearity and complex interactions between agricultural variables (Breiman, 2001 ; Belgiu & Drăguţ, 2016).
- **XGBoost (Extreme Gradient Boosting)**: The model was configured with a learning rate of 0.3, a maximum depth of 6, and no shrinkage rate. XGBoost was chosen for its speed of execution, built-in regularization, and superior performance in predicting large datasets (Chen & Guestrin, 2016 ; Mitchell & Frank, 2017).
- **KNN (K-Nearest Neighbors)** : The k-nearest neighbors model was implemented using Euclidean distance to measure similarity between observations. This approach is frequently used in precision agriculture for classification and regression, due to its simplicity and tolerance to noise in the data (Fix & Hodges, 1951 ; Cover & Hart, 1967).
- **Linear regression (LR)** : Used as *a baseline* model, it allows the performance of nonlinear models to be compared to a simple, explicit, and interpretable approach (Montgomery et al., 2012).
- **Artificial neural networks (ANN)** : A multilayer network was tested with three hidden layers, a ReLU activation function, a backpropagation algorithm for optimization, and a *dropout layer* (rate of 0.5) applied to each hidden layer. The model was trained over 50 epochs with a *batch size* of 32. ANNs are known for their ability to capture complex, nonlinear relationships between agroclimatic variables and yields (LeCun et al., 2015 ; Schmidhuber, 2015).
- **AdaBoost (Adaptive Boosting)** : This model was parameterized with 100 base estimators and a learning rate of 0.1. AdaBoost combines several weak classifiers to obtain a robust model, giving greater weight to errors from previous iterations (Freund & Schapire, 1997).
- **Stacking**: An ensemble approach was used, combining several base models (LR, RF, XGBoost, and ANN) with a linear meta-regression in the top layer. *Stacking* aims to exploit the complementarity between models to improve the overall performance of the prediction system (Wolpert, 1992).

– **Kernel Ridge Regression**: This hybrid model combines ridge regression and the *kernel trick* to model nonlinear relationships, while controlling overfitting through regularization (Saunders et al., 1998)

2.7. Performance evaluation

2.7.1. Cross-validation and evaluation of metrics

The models were evaluated using 5-fold cross-validation, ensuring class balance in each fold and a robust estimate of generalization error. This approach reduces the variance associated with data partitioning and provides a more stable and representative evaluation of the model's overall performance (James et al., 2013). Predictive performance was quantified using several statistical indicators commonly used for the evaluation of supervised models (Chicco & Jurman, 2020).

1. Recall (Sensitivity)

Recall measures the model's ability to correctly identify positive observations among all truly positive cases. It is defined by ;

$$Recall = \frac{TP}{TP+FN}$$

Where TP represents the number of true positives and FN the number of false negatives. A high value indicates that the model is good at detecting all positive cases

2. Precision

Precision reflects the proportion of correct positive predictions among all positive predictions made by the model:

Where FP corresponds to the number of false positives. A high accuracy indicates that the model produces few false alarms.

3. Specificity

Where TN refers to the number of true negatives. This metric complements recall by ensuring that the model does not confuse negative cases with positive ones.

4. Adjusted F1-score

The adjusted F1-score is the harmonic mean between precision and recall, providing a balanced overall measure of model performance:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Specificity evaluates the model's ability to correctly identify negative observations among all truly negative cases :

This metric is particularly useful when classes are imbalanced, as it penalizes models that favor a single
class.

5. Accuracy

Accuracy expresses the total proportion of correct predictions (positive and negative) across all observations :

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

However, in the case of unbalanced data, this metric can give a misleading impression of performance, as
it does not adequately reflect the model's ability to identify minority classes.

2.4.1. Performance indicators

Model performance was evaluated using classic, proven metrics commonly used in agricultural yield modeling. These indicators allow for the simultaneous assessment of the accuracy, robustness, and generalizability of the algorithms tested.

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) measures the square root of the mean of the squared differences between observed and predicted values. It particularly highlights the impact of large errors and outliers:

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

A lower RMSE value indicates a better match between observations and predictions (Willmott & Matsuura,
2005).

Coefficient of determination (R^2)

The **coefficient of determination (R^2)** expresses the proportion of the total variance of the observations

explained by the model:

$$R^2=\frac{\sum_{i=1}^{n}(yi-\widehat{yi})^2}{(yi-\overline{yi})^2}$$

An R^2 value close to 1 indicates a high fit of the model to the data (Nagelkerke, 1991)

Mean Absolute Percentage Error (MAPE)

The **Mean Absolute Percentage Error (MAPE)** measures the mean absolute error as a percentage of the
observed values:

$$MAPE=\frac{1}{n}\sum_{i\text{-}1}^{n}\left|\frac{yi-\widehat{yi}}{yi}\right|$$

This metric facilitates interpretation in an agricultural and economic context, as it directly expresses the
average relative error as a percentage (Kim & Kim, 2016)

Relative Root Mean Square Error (RRMSE)

The **Relative RMSE (RRMSE)** corresponds to the RMSE normalized by the mean of the observed values,
expressed as a percentage:

$$RRMSE=\frac{RMSE}{\overline{y}}$$

Lower values indicate better predictive power of the model.

Mean Bias Error (MBE)

Mean Bias Error (MBE) quantifies the model's average tendency to overestimate (MBE > 0) or underestimate (MBE < 0) observations :

$$MBE=\frac{\sum_i(\overline{yi}-yi)}{n}$$

This measure provides information on the presence of systematic bias in predictions.

Accuracy

Accuracy provides an integrative measure of performance, expressed as the complement of MAPE:

$$Accuracy=1\text{-}MAPE$$

It allows for a normalized comparison between models, regardless of the data scale.

Root Relative Squared Error (RRSE)

Root Relative Squared Error (RRSE) compares the quadratic errors of the model to those of a reference model based on the average of the observations :

$$RRSE=\sqrt{\frac{\sum_{i=1}^{n}(\widehat{yi}-yi)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2}}$$

A value close to zero indicates better performance of the model compared to the average model.

Relative Absolute Error (RAE)

Relative Absolute Error (RAE) evaluates the sum of the model's absolute errors compared to those of a reference model using the average of the observed values:

$$RAE=\frac{\sum_{i=1}^{n}|yi-\widehat{yi}|}{\sum_{i=1}^{n}|yi-\bar{y}|}$$

A value less than 1 indicates that the tested model predicts better than the average model.

Mean Squared Error (MSE)

Mean Squared Error (MSE) corresponds to the average of the squares of the differences between observed and predicted values:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(yi-\widehat{yi})^2$$

It is a direct measure of prediction error in quadratic units

Correlation coefficient (r)

The correlation coefficient (r) assesses the strength and direction of the linear relationship between observed and predicted values:

$$r=\frac{\sum_{i=1}^{n}(yi-\bar{y})\overline{(\widehat{yi}-\bar{\widehat{y}})}}{\sqrt{\sum_{i=1}^{n}(yi-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\widehat{yi}-\bar{\widehat{y}})^2}}$$

Values close to 1 indicate a strong positive correlation between observations and predictions.

Where:

- $\widehat{yi}$ represents the predicted value for sample i,
- yi represents the actual value for sample i,
- $\bar{yi}$ represents the mean of the actual values

These indicators, commonly used in agricultural modeling studies, allow for a joint assessment of the accuracy, robustness, and generalizability of the tested algorithms. Finally, an **accuracy function (Equation 1)** was used as an integrative metric to provide a standardized and comprehensive comparison of model performance.

2.5. Ethics Statement

This study was approved by the Cameroonian Ethical Committee for Research affiliated with the Ministry of Agriculture and Rural Development in accordance with the ethical guidelines and regulations set forth by the National Ethics Committee for Human and Agricultural Research. Verbal informed consent was obtained from all participants prior to the interviews.

3. Résults

3.1. Features importances

To evaluate the performance of nine different models for yield prediction based on agroeconomic data, an analysis of the importance of variables was performed using the Random Forest model. Figure 2 reveals that profit per hectare (PROFIT/Ha) overwhelmingly dominated with a relative importance score of 0.895, followed by quantities of fertilizer (0.034) and, to a much lesser extent, total area (0.010), production cost (0.008), and number of household members (0.006) . Based on their level of importance, 10 variables (including the 5 main ones and 5 additional ones with less impact) were selected for model development and comparative evaluation, while all others were discarded to optimize accuracy and generalization.

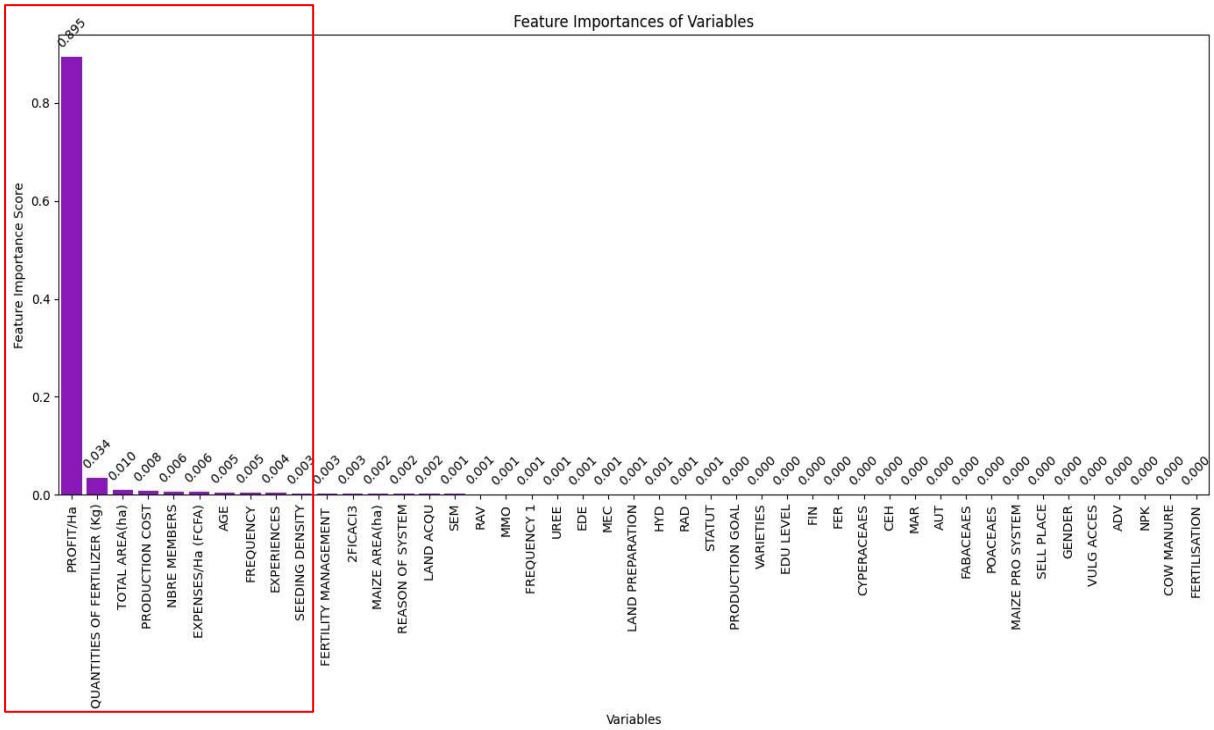


Figure 2. Features importances of variables.

3.2. Correlation Analysis of agro economic data and maize yield

The correlation matrix reveals key relationships between agro-economic variables and yield (Yield, Kg/ha), highlighting the most decisive drivers of agricultural performance. Yield is strongly correlated with profit/ha (r=0.85), reflecting a direct link between productivity and profitability. Expenses/ha also show a notable correlation with yield (r=0.67), suggesting that increased investment in inputs significantly improves productivity. Fertilizer use is relevant (r=0.45), confirming its central role in agricultural intensification. Production costs show a moderate but positive relationship with yield (r=0.48), while total area and producer age are weakly associated (r=0.11 and r=0.01 respectively), reflecting a limited influence. Experience correlates more strongly with age (r = 0.62) than with yield (r = 0.15), illustrating that expertise alone does not guarantee productivity gains without input support. Finally, planting density is virtually independent of other variables and yield (r = 0.01), emphasizing that technical management must go beyond simply the quantity of plants. These results highlight the strategic importance of optimizing inputs (fertilizer, targeted spending) to maximize yields and increase agricultural profitability in a context of limited resources.

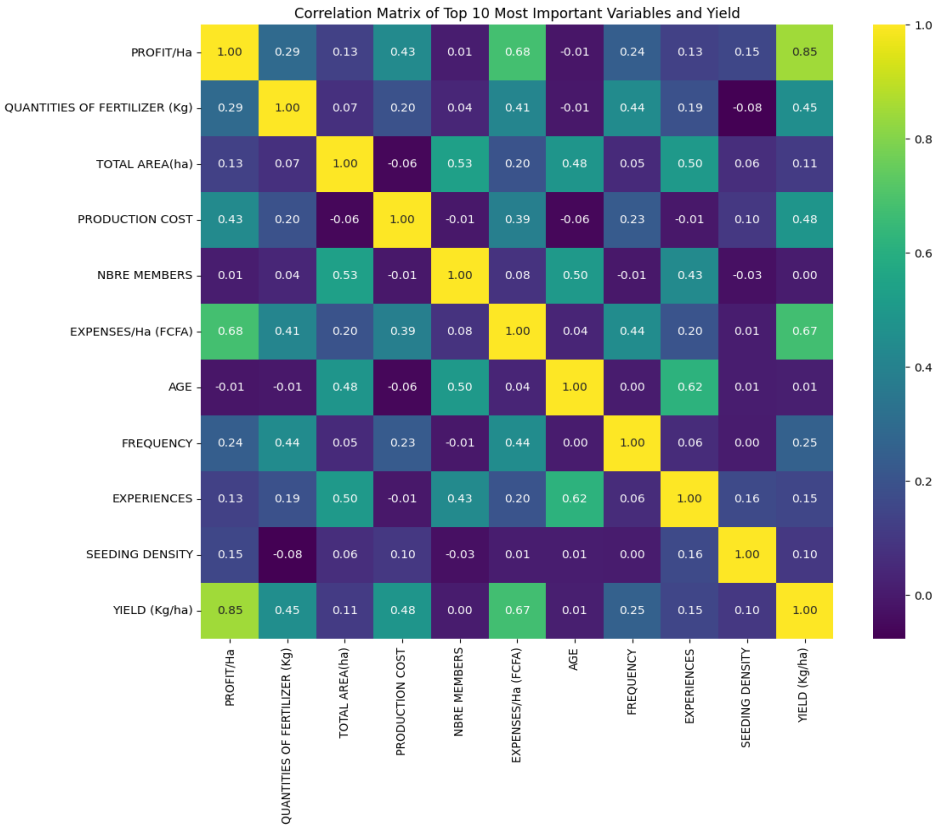


Figure 3. Pearson's correlation coefficient (r) between agro economic data and maize yield.

3.3. Évaluation des performances des modèles maize yield prediction

In this study, nine machine learning models were trained using observed yields and socioeconomic and agronomic variables. The performance of these models was evaluated using five-fold cross-validation, and the results were summarized according to the different approaches and time windows (Figure 4). Taking into account all evaluation indicators (R^2, RMSE, MAE, MSE, MAPE, RRMSE, and correlation), three models stand out clearly: XGBoost, Random Forest, and Stacking. These models show very high and almost similar accuracies, with R^2 values between 0.97 and 0.98, an MSE of 0.02 to 0.03, an RMSE of 0.12 to 0.16, an MAE of 0.06 to 0.09, and a MAPE of 17.32 to 34.72, reflecting remarkable robustness and excellent generalization ability. Although the other models (ANN, Adaboost, SVR, KR, RL, KNN) also perform well with an R^2 greater than 0.89, their errors remain higher (MSE: 0.09– 0.21; RMSE: 0.25–0.46; MAE: 0.19–0.22; MAPE: 52.12–107.14) (Figure 4). Overall, these results demonstrate the superiority of ensemble methods (XGBoost, Random Forest, Stacking) over linear approaches and proximity models, both in terms of absolute and relative accuracy. This performance can be explained by their ability to capture complex nonlinear relationships and better manage inter-plot variability. Thus, from a precision agriculture decision support perspective, XGBoost appears to be the most stable and effective solution, closely followed by Stacking. The use of these models could contribute to

significantly improving corn yield forecasting and, consequently, strengthening agricultural planning in Cameroon.

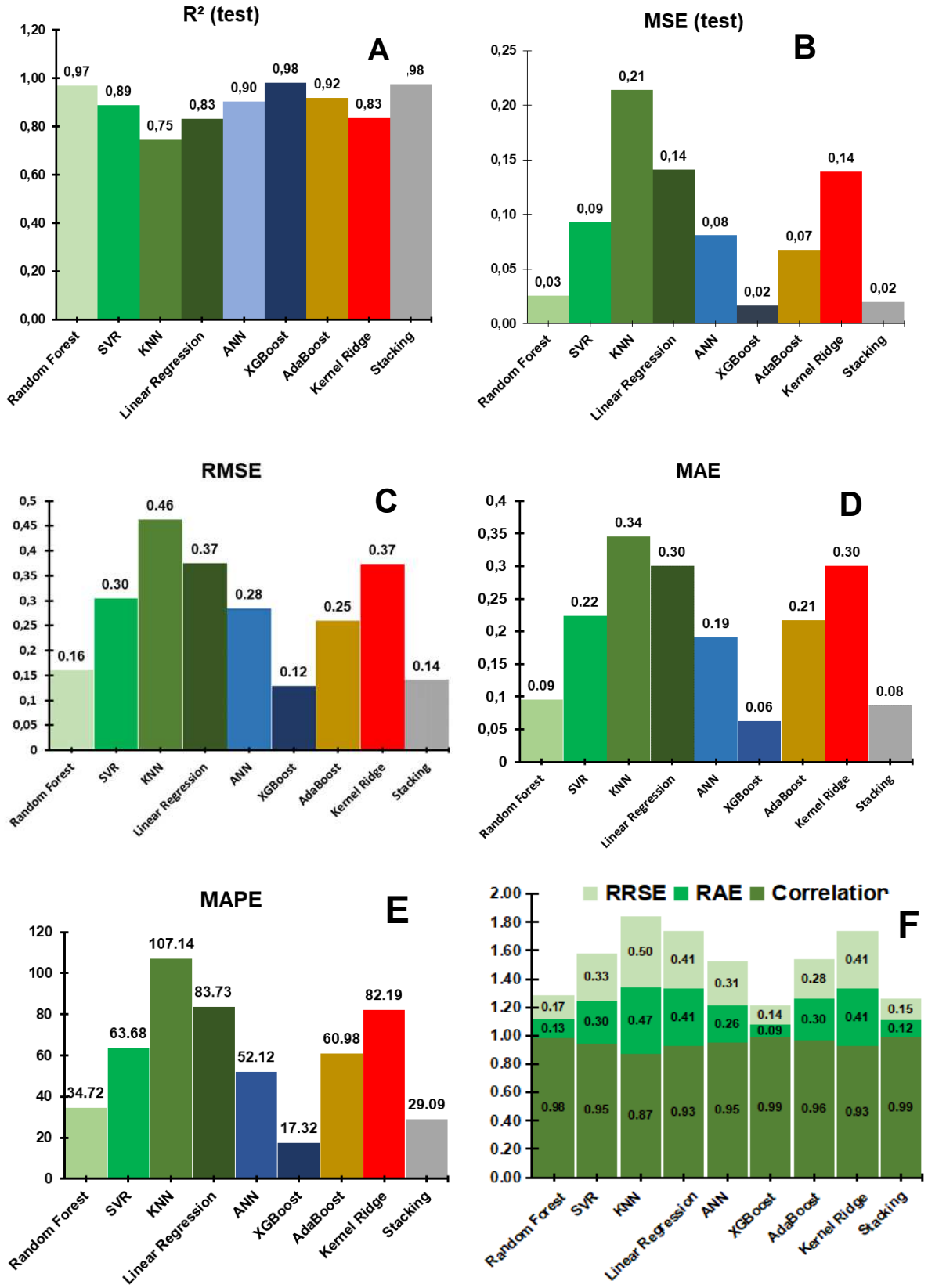


Figure 4. Comparison of the predictive skill of the proposed maize yield prediction models in terms of the R2 ; MAE ; RMSE ; MAPE ; MSE ; RRSE ;RAE and Correlation r

554

555 **3.4. Maize yield predictions**

556 Based on the trained models of RF, XGBoost, and stacking, maize yields of 354 farmers were predicted. The
557 residuals of the prediction results of these models all passed the Shapiro Wilk test and obeyed normal
558 distribution, which showed that these regression models were acceptable (Figure 6). The scatter diagrams of
559 the predicted and observed yields of all the models are shown in Figure 4. We found that the predicted and
560 observed yields showed a good linear fit with R(2) of about 0.98 0.97. Such results indicated that the three
561 machine learning models can predict the yield of maize at the county level with higher accuracy, in the order
562 is XGBoost > stacking > RF. Although all predicted yields were much closer to the 1:1 line, consistent
563 underestimations were found for all models and all time windows. Moreover, the prediction areas were
564 overestimated for low yields observed with smaller deviations, while underestimated for high observed
565 yields with relatively greater deviations. Nevertheless, all errors of thepredicted yields are within 10%,
566 suggesting that RF, GPR, and SVM models perform well for crop prediction at larger scales.

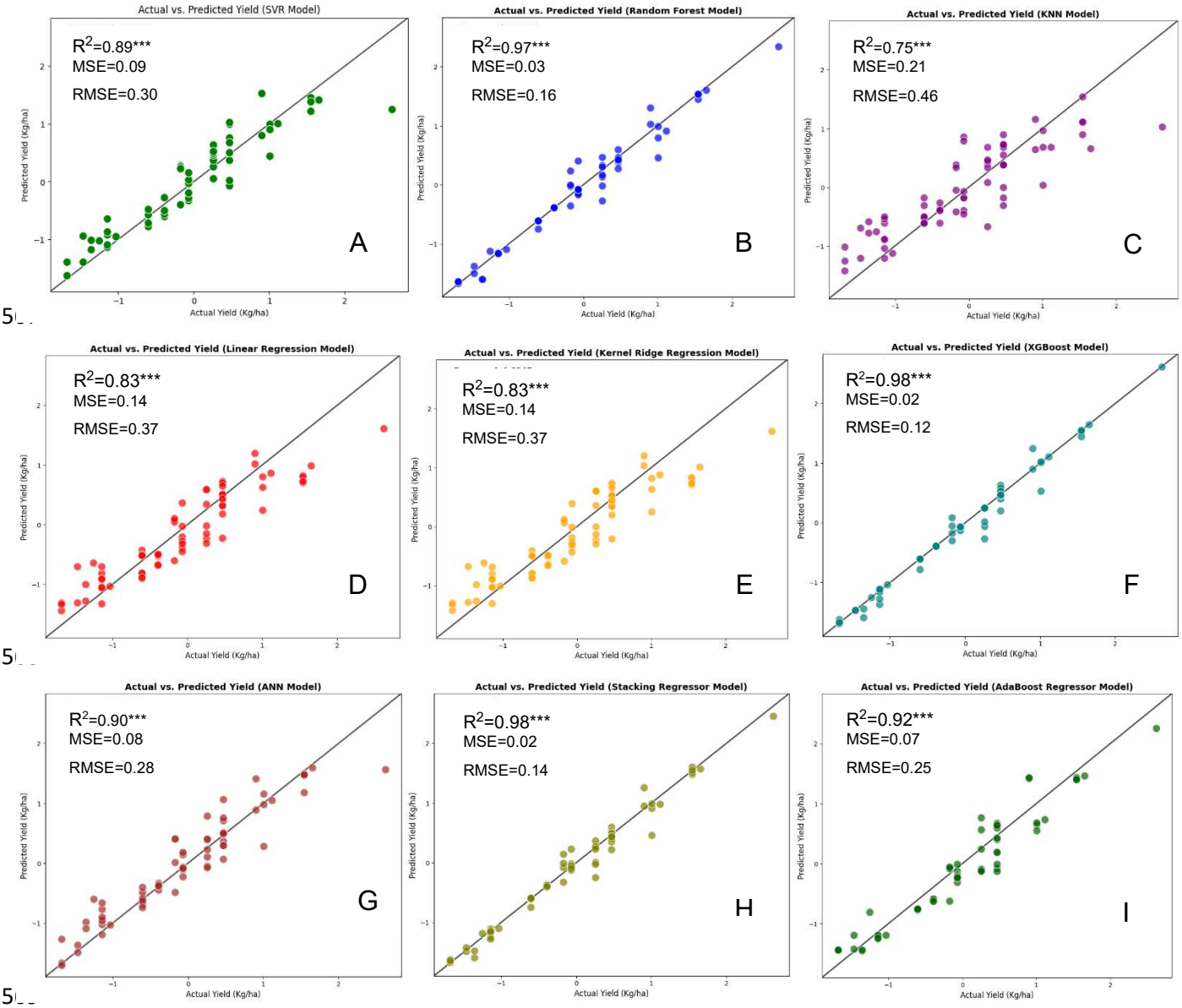


Figure 5. Scatter plots of observed yield and predicted yield maize. Multiple linear regression (MLR), ridge regression (RR), k-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), artificial neural networks (ANN), and extreme gradient boosting (XGBoost) Adaboost, staking.

3.5. Cross-validation of models

Table 1 shows the prediction performance of the different models based on a k-fold cross-validation test set at 5 folds. The results confirm the differential robustness of the approaches tested in predicting maize yield. The ensemble methods stand out clearly: Random Forest (R^2 = 0.9396 ±0.0141; RMSE = 0.242 ± 0.033) and Stacking (R^2 = 0.9388 ± 0.0154; RMSE = 0.242 ± 0.033) show remarkable stability, characterized by low variance around the mean, suggesting excellent generalization ability. Similarly, XGBoost performs very similarly (R^2 = 0.9365 ± 0.0255; RMSE = 0.243± 0.054), confirming its superiority already observed in the initial evaluation and its ability to effectively model complex nonlinear relationships. In an intermediate position, AdaBoost (R^2 = 0.9045 ± 0.0202; RMSE = 0.303 ± 0.026) offers

an acceptable compromise between accuracy and stability. Conversely, simpler models such as linear regression (R^2 = 0.7533 ± 0.0878; RMSE = 0.483 ± 0.070) and Kernel Ridge (R^2 = 0.7552 ± 0.0871; RMSE = 0.481 ± 0.070) prove to be more modest, with increased variability, reflecting limited reliability in an operational context. KNN recorded the lowest results (R^2 = 0.7358 ± 0.0472; RMSE = 0.508 ± 0.066), confirming its sensitivity to sample fluctuations. SVR, although superior to linear approaches (R^2 = 0.8235 ± 0.0258; RMSE = 0.418 ± 0.065), still lags behind ensemble models. Overall, these results validate the hypothesis that ensemble algorithms (Random Forest, XGBoost, Stacking) consistently outperform linear approaches and proximity-based models. Their low standard deviation attests to their increased reliability and robustness across different subsamples, reinforcing their relevance for large-scale corn yield prediction.

Table 1. Comparative performance of corn yield prediction models obtained by cross-validation (Mean ± Standard Deviation, k-fold CV).

	Cross-validation (Mean ± Std Dev)		
Models	**R-squared Scores**	**MSE Scores**	**RMSE Scores**
Random Forest	0.9396 ± 0.0141	0.0595 ± 0.0170	0.2417 ± 0.0328
SVR	0.8235 ± 0.0258	0.1788 ± 0.0555	0.4178 ± 0.0647
KNN	0.7358 ± 0.0472	0.2628 ± 0.0669	0.5084 ± 0.0663
Linear Regression	0.7533 ± 0.0878	0.2378 ± 0.0716	0.4827 ± 0.0699
ANN	-	-	-
XGBoost	0.9365 ± 0.0255	0.0621 ± 0.0293	0.2433 ± 0.0541
AdaBoost	0.9045 ± 0.0202	0.0926 ± 0.0159	0.3031 ± 0.0263
Kernel Ridge	0.7552 ± 0.0871	0.2364 ± 0.0717	0.4811 ± 0.0703
Stacking	0.9388 ± 0.0154	0.0597 ± 0.0172	0.2422 ± 0.0325

3.6. Error analysis

Figure 5 illustrates the distribution of residuals (actual values – predicted values) obtained for each of the regression models tested on the validation set. From a statistical perspective, a high-performance model is characterized by residuals centered around zero and reduced dispersion, reflecting low error variance and the absence of systematic bias. This dispersion is generally represented by a narrow interquartile range and short whiskers on the boxplot. In this study, the XGBoost Regressor and Stacking Regressor models have the most compact residual distributions, indicating high predictive accuracy and excellent stability of their estimates. The Random Forest Regressor also shows a marked concentration of residuals around zero, suggesting comparable performance, particularly in terms of minimizing mean squared errors (MSE) and relative squared errors (RSE). To a slightly lesser extent,

the AdaBoost Regressor shows moderate dispersion, reflecting an acceptable trade-off between bias and variance. In contrast, the SVR, Kernel Ridge Regressor, and Linear Regression models have more spread-out residual distributions, revealing increased variability in prediction errors. It should be noted that linear regression maintains satisfactory symmetry around zero, demonstrating a good balance between overestimation and underestimation of observed values. Finally, the KNN Regressor and Artificial Neural Network (ANN) models show the greatest dispersion, accompanied by multiple outliers, reflecting instability in model generalization and high sensitivity to data noise. The behavior of the ANN model, characterized by a wide box and long whiskers, suggests overfitting or suboptimal hyperparameter configuration, consistent with its lower coefficient of determination (R^2) and high error metrics.

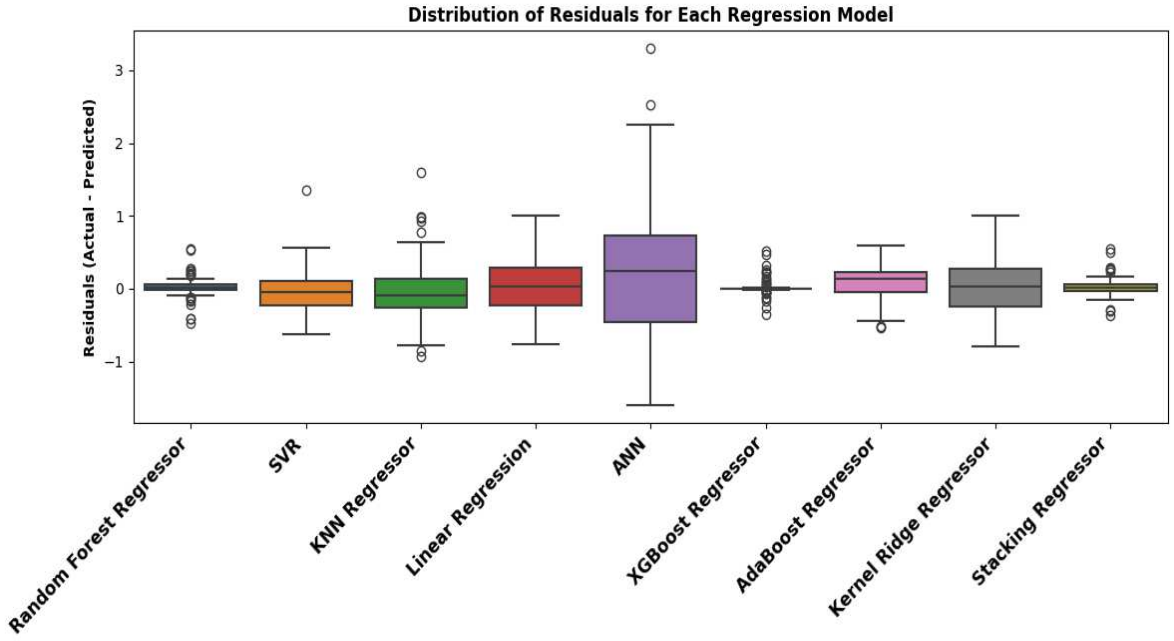


Figure 6. Box plot Comparison of Residual Distributions.

3.7. Analysis of residual distributions

Figure 7 illustrates the distribution of residuals for the XGBoost and Stacking Regressor models. For XGBoost, the residuals are highly concentrated around zero with a quasi-symmetrical distribution and relatively low variance. The majority of the differences between observed and predicted values fall within the narrow range [−0.1; +0.1], reflecting high accuracy and the absence of marked systematic bias. However, the presence of a few extreme values (positive and negative) suggests that the model is somewhat sensitive to atypical observations or variations not captured by the explanatory variables (Figure 7a). The Stacking Regressor also has a distribution centered around zero, confirming the absence of bias, but with a slightly wider dispersion than that of XGBoost. The density curve shows

thicker tails, reflecting increased variability in the estimation of extreme values. This behavior reflects the hybrid nature of the model, which combines several algorithms and can lead to greater heterogeneity in predictions (Figure 7b). Overall, both models show good generalization ability, but XGBoost appears to be the most accurate and robust, with more concentrated residuals and better predictive reliability. The Stacking Regressor, although effective, remains slightly more exposed to variance, which corroborates the performance figures obtained previously (absolute and quadratic errors higher than those of XGBoost).

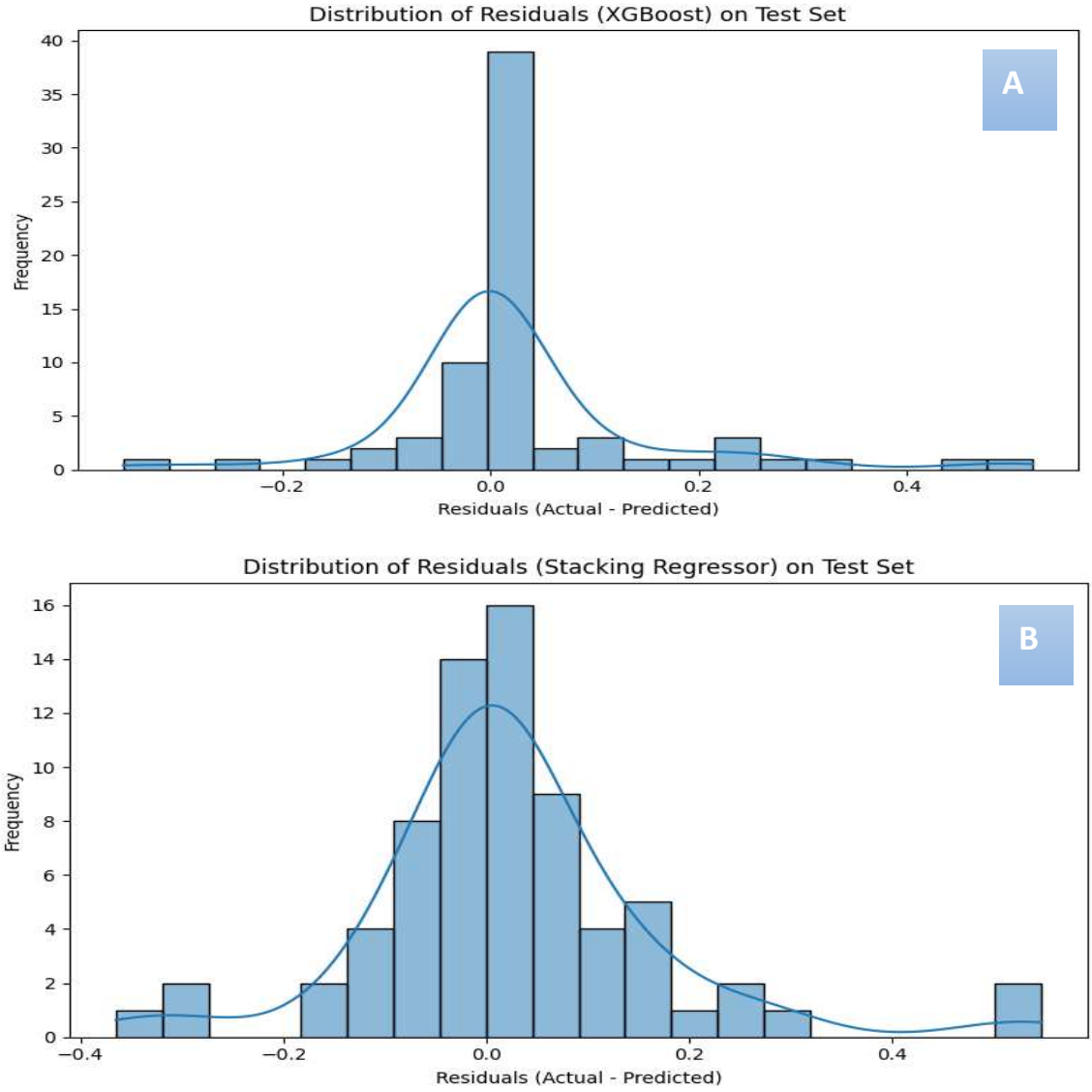


Figure 7. Distribution of residuals on test set.

3.8. Learning curve analysis

Figure 9 shows the learning curves for the different models evaluated. There are clear differences between them. The XGBoost Regressor and Random Forest Regressor models show high and stable training scores, while their cross-validation scores gradually improve as the sample size increases, until they approach training performance. This reflects good generalization ability and a low risk of

overfitting, particularly when more data is available. The Stacking Regressor exhibits similar behavior, confirming its effectiveness and robustness. The AdaBoost Regressor also achieves good scores, but with a slightly wider gap between training and validation, suggesting greater sensitivity to data variance. The SVR and Kernel Ridge Regressor models show higher training scores than validation scores, with a tendency to overfit on small datasets. However, this gap decreases as the number of examples increases, reflecting a gradual improvement in generalization. Linear regression is characterized by rapidly converging curves, with higher bias but little variance, which is typical of simple models with low risk of overfitting but limited performance ceilings. Finally, the KNN Regressor is distinguished by a significant and persistent gap between training and validation scores, even with larger datasets, revealing marked overfitting and poor generalization ability under the experimental conditions of this study.

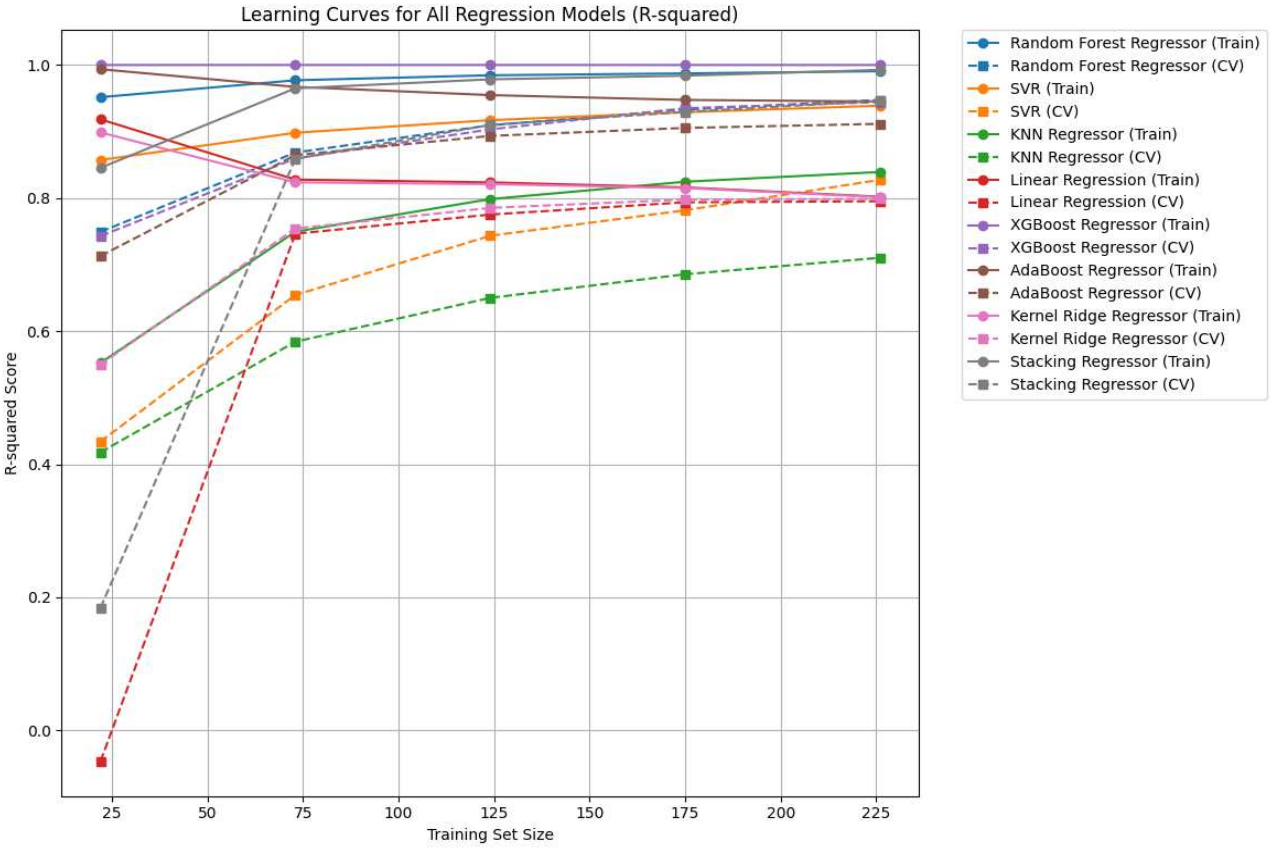


Figure 9. Learning curves of differents models use in the study.

4. Discussion

Our results reveal that technical and economic management of inputs has a much greater influence on corn yield than sociodemographic characteristics. Variables such as profit, fertilizer use, and cultivated area appear to be major factors, with importance values ranging from 0.895 to 0.034. This positive

relationship is confirmed by the correlations observed (r = 0.85 to r <0.20), indicating that an increase in these inputs systematically results in improved yields.

From a methodological perspective, ensemble methods proved superior to all other approaches tested. The XGBoost model performed exceptionally well (R^2 = 0.98), closely followed by Random Forest (R^2 = 0.97) and Stacking (R^2 = 0.97). Conversely, linear (regression, Ridge) and proximity (KNN) models showed modest performance (R^2 < 0.84; MAPE > 80%), while intermediate models (ANN, Adaboost, SVR) showed acceptable but less stable results (R^2 ~ 0.90). These results are consistent with trends observed in recent studies. For example, Li et al. (2025) in China showed that XGBoost applied to multi-source data (vegetation indices, soil and geographic data) explained up to 85% of the variance in winter wheat yield, outperforming Lasso regression and random forest. Similarly, a study conducted in Ghana found that Random Forest could explain approximately 81% of the variance in maize yield, confirming the importance of soil variables (Asare et al., 2023). In another context, Chen et al. (2025) demonstrated that XGBoost improved the robustness of corn yield predictions by capturing complex nonlinear relationships, while linear approaches lost accuracy. Finally, work in South Africa using unmanned aerial vehicle (UAV) s showed that RF and XGBoost could achieve an R^2 of 0.95 depending on the growth stage of the corn, confirming their applicability in different agricultural environments (Masocha et al., 2023). This performance hierarchy (ensemble methods > neural networks/SVR > simple models) can be explained by the ability of ensemble algorithms to model complex, nonlinear interactions between variables, unlike linear models, which assume proportional relationships that are often unsuited to agricultural phenomena.

The remarkable performance of XGBoost could also be explained by its gradient boosting optimization structure, which iteratively refines predictions and weights decision trees according to their effectiveness. Its objective function incorporates a regularization term (L2), limiting overfitting while capturing complex relationships, such as the impact of access to extension services on yield. In comparison, KNN proved to be very sensitive to sample fluctuations, reflecting a high risk of overfitting and low generalization ability.

A key finding of this research is that accessible and inexpensive socioeconomic data can compete with the climate and satellite data that are often used. This provides significant added value in low-resource agricultural systems, particularly in sub-Saharan Africa, where access to high resolution remote sensing data remains limited. Thus, the use of ensemble models applied to economic and social data is a promising alternative for yield forecasting and decision support. Nevertheless, certain limitations should be mentioned. First, the sample size (354 producers) remains relatively small, which may limit the generalizability of the results. Second, the data used are static and self-reported, which may introduce biases related to producers' memory or reporting. Finally, the absence of climatic and spectral data prevents direct comparison with multi-source models. Despite these limitations, the observed

robustness of ensemble methods suggests that these models are not only reliable but also highly transferable to other agricultural contexts.

5. Conclusion

This study demonstrates the strong potential of machine learning algorithms, particularly ensemble methods, for robustly predicting maize yield based on socio-agronomic and economic data collected from 354 producers in Vina, Cameroon. The evaluation clearly indicates the high predictive power of XGBoost (R^2=0.98; RMSE = 0.128) and Random Forest (R^2=0.97), closely followed by the hybrid Stacking approach (R^2=0.976). These tree-based models substantially outperform linear and proximity models such as Multiple Regression and K-Nearest Neighbors (KNN). Analysis of the variables suggests that profit per hectare and fertilizer use are major determinants of yield, while the socio-demographic characteristics of producers had a marginal impact. These findings underscore that optimizing agricultural inputs remains the key driver of productivity in this specific context of limited resources. From a scientific standpoint, this work makes an original contribution by suggesting that socioeconomic data, often neglected in favor of spectral or climatic data, can be highly effective in producing reliable predictions when appropriate ML models are employed. From a practical standpoint, the models developed offer decision-makers a highly effective tool for planning and providing targeted technical support to producers in the Vina region.

Study Limitations and Future Research

The main limitations of this work are twofold: first, the use of uncalibrated empirical data. Specifically, the corn yields were obtained using locally available, non-calibrated measurement instruments, which may introduce a significant degree of measurement error or estimation bias. Second, the static and geographically limited nature of the dataset (collected in one campaign in Vina) necessitates caution in the broad generalization of the findings. These limitations open up clear prospects for future research integrating time series, multi-source data (remote sensing, soil data), and deep learning approaches to enhance the robustness and generalizability of the models across different agro-ecological zones.

Ethics approval and consent to participate; The research protocol was reviewed and granted ethical approval by the **Cameroonian Ethical Comity for Research**, affiliated with the **Ministry of agriculture and rural development**. Prior to their participation, verbal informed consent was obtained from each voluntary interviewee by the research team. Interviewees were provided with detailed information about the study, and face-to-face interviews were conducted to ensure comprehension. The research team, comprising personnel from the Ministry of Agriculture and Rural Development, provided all necessary explanations and clarifications.

Funding. No funding was received for conducting this study.

Consent to Publish declaration: not applicable.

Authors' contributions. SKNG and BAF conceptualized the paper, collected data, SKNG analyzed the data, wrote the first draft and the final version. DTG and KMJP were involved in the study design, the critical review of all drafts and approval of the final version. All authors read and approved the final manuscript.

Data availability All data generated and analyzed during this study are included in the manuscript and its supplementary files.

Ethics and Consent to Participate declarations: not applicable.

References

Alexandrov, V., & Hoogenboom, G. (2000). The impact of climate variability and change on crop yield in Bulgaria. *Agricultural and Forest Meteorology, 104*, 315–327.

Arvor, D., Bégué, A., Dubreuil, V., & Nelson, A. (2023). Global maize yield prediction using machine learning approaches: Evidence from 37 developing countries. *arXiv*. https://arxiv.org/abs/2312.02254

Asare, E., Aidoo, O. F., & Boateng, E. (2023). Application of random forest for maize yield prediction under varying soil and climatic conditions in Ghana. *Frontiers in Sustainable Food Systems, 7*, 11403005. https://doi.org/10.3389/fsufs.2023.11403005

Aschonitis, V., Mastrocicco, M., Colombani, N., Salemi, E., Kazakis, N., Voudouris, K., & Castaldelli, G. (2012). Assessment of the intrinsic vulnerability of agricultural land to water and nitrogen losses via deterministic approach and regression analysis. *Water, Air, and Soil Pollution, 223*, 1605–1614.

Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Apress. https://doi.org/10.1007/978-1-4302-5990-9

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

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., et al. (2017). Crop yield predictions: High-resolution statistical model for intra-season forecasts applied to corn in the U.S. *Gro Intelligence, Inc.*

Challinor, A. J., Watson, J., Lobell, D. B., Howden, S., Smith, D., & Chhetri, N. (2014). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change, 4*, 287–291.

Chen, L., Wang, J., Sun, Y., & Huang, Z. (2025). UAV-based multispectral imagery and machine learning for maize yield prediction. *Computers and Electronics in Agriculture, 224*, 109050. https://doi.org/10.1016/j.compag.2025.109050

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

Chen, X., Li, J., & Wang, Y. (2025). Machine learning-based prediction of maize yield using socioeconomic and climatic data. *Agricultural Systems, 209*, 103721. https://doi.org/10.1016/j.agsy.2025.103721

Cheng, E., Zhang, B., Peng, D., Zhong, L., Yu, L., Liu, Y., et al. (2022). Wheat yield estimation using remote sensing data based on machine learning approaches. *Frontiers in Plant Science, 13*, 1090970. https://doi.org/10.3389/fpls.2022.1090970

Cheng, E., Zhang, B., Peng, D., Zhong, L., Yu, L., Liu, Y., Xiao, C., Li, C., Li, X., Chen, Y., Ye, H., Wang, H., Yu, R., Hu, J., & Yang, S. (2022). Wheat yield estimation using remote sensing data based on machine learning approaches. Frontiers in Plant Science, 13, 1090970. https://doi.org/10.3389/fpls.2022.1090970

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture, 151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters, 13*(11), 114003. https://doi.org/10.1088/1748-9326/aae159

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (pp. 1–15). Springer.

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

Fan, J., Chen, X., & Zhang, Y. (2023). Machine learning-based crop yield prediction: A comparative study of XGBoost, Random Forest, and deep learning models. *Computers and Electronics in Agriculture, 210*, 107999. https://doi.org/10.1016/j.compag.2023.107999

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: Nonparametric discrimination—Consistency properties* (Technical Report No. 4). USAF School of Aviation Medicine.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., & Müller, J. (2013). Random Forests modeling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management, 116*, 142–150. https://doi.org/10.1016/j.agwat.2012.07.003

Gao, J., Zhang, Y., Feng, P., Liu, Y., & Li, X. (2023). Maize yield prediction with machine learning, spectral variables, and irrigation management. *Computers and Electronics in Agriculture, 205*, 107624. https://doi.org/10.1016/j.compag.2023.107624

González Sánchez, A., Frausto Solís, J., & Ojeda Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research, 12*(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439

Habib-ur-Rahman, M., Ahmad, A., Raza, A., Hasnain, M. U., Alharby, H. F., Alzahrani, Y. M., Bamagoos, A. A., Hakeem, K. R., Ahmad, S., Nasim, W., Ali, S., Mansour, F., & El Sabagh, A. (2022). Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia. *Frontiers in Plant Science*, *13*, 925548. https://doi.org/10.3389/fpls.2022.925548

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks, 5*(6), 989–993. https://doi.org/10.1109/72.329697

Han, L., Yang, G., Feng, H., Zhou, C., & Yang, H. (2020). Hyperspectral-based prediction of wheat yield using machine learning techniques. *Remote Sensing, 12*(9), 1–20.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hoogenboom, G., White, J. W., & Messina, C. D. (2004). From genome to crop: Integration through simulation modeling. *Field Crops Research, 90*(1), 145–163. https://doi.org/10.1016/j.fcr.2004.07.014

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Iizumi, T., & Ramankutty, N. (2015). How do weather and climate influence cropping area and intensity? *Global Food Security*, *4*, 46–50.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE, 11*(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

Karimi, Y., Prasher, S., Madani, A., & Kim, S. (2008). Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. *Canadian Biosystems Engineering, 50*(7), 13–20.

Kharal, A. S., Mahar, S. A., Mushtaque, M. I., Magsi, A., & Mahar, J. A. (2024). A model for wheat yield prediction to reduce the effect of climate change using support vector regression. *VFAST Transactions on Software Engineering, 12*(2), 192–212. https://doi.org/10.21015/vtse.v12i2.1855

838 Kharal, A. S., Mahar, S. A., Mushtaque, M. I., Magsi, A., & Mahar, J. A. (2024). A Model for Wheat Yield
839 Prediction to Reduce the Effect of Climate Change Using Support Vector Regression. VFAST
840 Transactions on Software Engineering, 12(2), 192–212. https://doi.org/10.21015/vtse.v12i2.1855

841 Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts.
842 *International Journal of Forecasting, 32*(3), 669–679.

843 Kingsley, J., Afu, S. M., Isong, I. A., Chapman, P. A., Kebonye, N. M., & Ayito, E. O. (2021). Estimation
844 of soil organic carbon distribution by geostatistical and deterministic interpolation methods: A case
845 study of the southeastern soils of Nigeria. *Environmental Engineering and Management Journal, 20*, 1077–
846 1085.

847 Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., Mushtaq, S., & Nguyen, V. P. (2018). Artificial
848 intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Computers*
849 *and Electronics in Agriculture, 155*, 324–338.

850 Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

851 LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
852 https://doi.org/10.1038/nature14539

853 Li, B., Yang, W., & Li, X. (2018). Application of combined model with DGM(1,1) and linear regression
854 in grain yield prediction. *Grey Systems: Theory and Application, 8*(1), 25–34. https://doi.org/10.1108/GS-
855 07-2017-0020

856 Li, H., Wang, J., & Zhao, Y. (2023). Comparative performance of machine learning algorithms for crop
857 yield prediction: Evidence from support vector regression and ensemble methods. *Agricultural Systems,*
858 *207*, 103613. https://doi.org/10.1016/j.agsy.2023.103613

859 Li, T., Zhou, Y., Li, X., Wu, J., & He, T. (2019). Forecasting daily crude oil prices using improved
860 CEEMDAN and ridge regression-based predictors. Energies, 12(19), Article 3603.
861 https://doi.org/10.3390/en12193603

862 Li, Y., Zhang, H., & Liu, Q. (2025). Application of XGBoost model and multi-source data for winter
863 wheat yield prediction in Henan Province of China. *Computers and Electronics in Agriculture, 215*, 108528.
864 https://doi.org/10.1016/j.compag.2025.108528

Lischeid, G., Webber, H., Sommer, M., Nendel, C., & Ewert, F. (2022). Machine learning in crop yield modeling: A powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology, 312*, 108698. https://doi.org/10.1016/j.agrformet.2021.108698

Maseko, S., Van Der Laan, M., Tesfamariam, E. H., Delport, M., & Otterman, H. (2024). Evaluating machine learning models and identifying key factors influencing spatial maize yield predictions in data intensive farm management. *European Journal of Agronomy*, 157, 127193. https://doi.org/10.1016/j.eja.2024.127193

Masocha, M., Mutanga, O., & Sibanda, M. (2023). Integrating UAV imagery and machine learning for maize yield prediction across growth stages in South Africa. *Remote Sensing Applications: Society and Environment, 29*, 100982. https://doi.org/10.1016/j.rsase.2023.100982

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends, 1*(2), 140–147. https://doi.org/10.38094/jastt1457

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends,* 1(2), 140–147. https://doi.org/10.38094/jastt1457

Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture, 139*, 103–114. https://doi.org/10.1016/j.compag.2017.05.002

Miller, T., Rahman, A., & Ghosh, S. (2023). Neural networks for crop yield prediction: A comparative analysis with machine learning models. *Computers and Electronics in Agriculture, 208*, 108012. https://doi.org/10.1016/j.compag.2023.108012

Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science, 3*, e127. https://doi.org/10.7717/peerj-cs.127

Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., & Gocić, M. (2015). Extreme learning machine based prediction of daily dew point temperature. *Computers and Electronics in Agriculture, 117*, 214–225. https://doi.org/10.1016/j.compag.2015.08.008

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In O. A. Montesinos López, A. Montesinos López, & J. Crossa,

Multivariate Statistical Machine Learning Methods for Genomic Prediction (pp. 109–139). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-89010-0_4

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., et al. (2016). Machine learning-based prediction of soil total nitrogen, organic carbon, and moisture content using VIS-NIR spectroscopy. *Biosystems Engineering, 152*, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018

Mutanga, O., Adam, E., & Cho, M. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation, 18*, 399–406. https://doi.org/10.1016/j.jag.2012.03.012

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692.

Naik, J., Satapathy, P., & Dash, P. (2018). Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Applied Soft Computing*, 70, 1167–1188. https://doi.org/10.1016/j.asoc.2018.06.008

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

Pham, H., & Olafsson, S. (2019a). Bagged ensembles with tunable parameters. *Computational Intelligence, 35*(1), 184–203. https://doi.org/10.1111/coin.12198

Pham, H., & Olafsson, S. (2019b). On Cesaro averages for weighted trees in the random forest. *Journal of Classification, 1–14*. https://doi.org/10.1007/s00357-019-09322-8

Romeijn, H., Faggian, R., Diogo, V., & Sposito, V. (2016). Evaluation of deterministic and complex analytical hierarchy process methods for agricultural land suitability analysis in a changing climate. *ISPRS International Journal of Geo-Information, 5*, 99.

Rosenberg, N. J. (1992). Adaptation of agriculture to climate change. *Climatic Change, 21*, 385–405.

Ruane, A. C., Major, D. C., Winston, H. Y., Alam, M., Hussain, S. G., Khan, A. S., Hassan, A., Al Hossain, B. M. T., Goldberg, R., & Horton, R. M. (2013). Multi-factor impact analysis of agricultural production in Bangladesh with climate change. *Global Environmental Change, 23*, 338–350.

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., et al. (2018). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment, 644*, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning* (pp. 515–521). Morgan Kaufmann.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

Shahhosseini, M., Hu, G., & Pham, H. (2019a). Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *arXiv:1908.05287*.

Shahhosseini, M., Hu, G., & Pham, H. (2019b). Optimizing ensemble weights for machine learning models: A case study for housing price prediction. In H. Yang, R. Qiu, & W. Chen (Eds.), *Smart service systems, operations management, and analytics* (pp. 1–14). Springer. https://doi.org/10.1007/978-3-030-30967-1_9

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., & Ebrahimie, E. (2014). Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: A new avenue in intelligent agriculture. *PLoS ONE, 9*(5), e97288. https://doi.org/10.1371/journal.pone.0097288

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222. https://doi.org/10.1023/B\:STCO.0000035301.49549.88

Subedi, B., Poudel, A., & Aryal, S. (2023). The impact of climate change on insect pest biology and ecology: Implications for pest management strategies, crop production, and food security. *Journal of Agriculture and Food Research*, *14*, 100733. https://doi.org/10.1016/j.jafr.2023.100733

Tadesse, T., Demisse, G. B., Zaitchik, B., et al. (2018). Building resilience to food insecurity in data-scarce regions. *Agricultural and Forest Meteorology, 262*, 402–413.

Tao, F., Li, Y., Wei, Y., Zhang, C., & Zuo, Y. (2025). Data–model Fusion Methods and Applications toward Smart Manufacturing and Digital Engineering. *Engineering*, S2095809925000244. https://doi.org/10.1016/j.eng.2024.12.034

Thornton, P. K., Jones, P. G., Alagarswamy, G., & Andresen, J. (2009). Spatial variation of crop yield response to climate change in East Africa. *Global Environmental Change, 19*, 54–65.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G. A., et al. (2011). Application of a Random Forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecological Modelling, 222*(8), 1471–1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yang, Q., Li, P., & Sun, J. (2024). Deep learning approaches for crop yield prediction: Challenges and opportunities. *Agricultural Systems, 212*, 103754. https://doi.org/10.1016/j.agsy.2024.103754

Yang, S., Li, L., Fei, S., Yang, M., Tao, Z., Meng, Y., & Xiao, Y. (2024). Wheat yield prediction using machine learning method based on UAV remote sensing data. *Drones, 8*(7), 284. https://doi.org/10.3390/drones8070284

Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhao, L., Wang, H., & Li, P. (2024). Enhancing crop yield prediction with XGBoost and remote sensing data: Evidence from maize and wheat systems. *Agricultural Systems, 212*, 103753. https://doi.org/10.1016/j.agsy.2024.103753

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.

🚀 **PLAN DE RÉDACTION POUR FAIRE TREMBLER LA TERRE**

Target : Agricultural Systems (IF ~6.6) / Computers & Electronics in Agriculture (IF ~8.3) / PLOS ONE thématique

---

TITRE REFORGE

"**Socioeconomic data alone can reliably predict smallholder maize yields: evidence from machine learning benchmarking in Cameroon's Adamawa region**"

ou version encore plus forte :

"**Beyond remote sensing: ensemble machine learning with farm survey data achieves high-accuracy maize yield prediction in data-scarce Sub-Saharan Africa**"

PLAN ARTICLE NIVEAU TOP REVUE

ABSTRACT (restructuré selon le template IMRAD strict)

Background/Problem — Précision agricole en ASS bloquée par le coût des données satellitaires/climatiques. Alternative socio-économique sous-explorée.

Objective — Tester si les données d'enquête agriculteurs (N=354, Vina, Cameroun) suffisent pour des prédictions fiables via ML.

Methods — Benchmark de 9 algorithmes ML avec cross-validation robuste k=5, analyse d'importance des variables, et évaluation de la robustesse par simulation de bruit de mesure.

Key finding — XGBoost (R^2=0.93 en CV, RMSE=0.243) et RF surpassent toutes les baselines avec **uniquement des données d'enquête**. Profit/ha, fertilisation et superficie sont les déterminants dominants.

Implication — Les systèmes de prédiction de rendement low-cost basés sur les données socio-économiques sont viables pour la planification agricole en ASS.

---

1. INTRODUCTION (totalement réécrite — 800 mots max, très ciblée)

Paragraphe 1 — Le problème mondial de la prédiction de rendement en ASS (2-3 citations récentes 2022-2025)

La sécurité alimentaire en ASS dépend de prédictions de rendement fiables. Or les approches dominantes (télédétection, modèles climatiques) restent inaccessibles aux petits États à faibles ressources.

Paragraphe 2 — Le paradoxe des données agricoles africaines

Des millions d'enquêtes agriculteurs existent (FAO, ministères, ONG) mais sont sous-exploitées pour la modélisation ML. Ces données socio-économiques ont-elles un pouvoir prédictif comparable aux données spectrales ?

Paragraphe 3 — État de l'art en ML agricole (ciblé)

Citer 6-8 études clés. Montrer que **personne ne s'est focalisé sur les données d'enquête seules en ASS pour le maïs** → gap clair.

Paragraphe 4 — La question de recherche (explicite)

"This study addresses three questions: (1) Can socioeconomic farm survey data alone achieve reliable maize yield prediction? (2) Which ML architectures best handle the noise and heterogeneity of self-reported data? (3) What socioeconomic factors most strongly predict yield in resource-limited smallholder systems?"

Paragraphe 5 — Contribution originale

Lister en 4 points la nouveauté (contexte africain + données enquête + benchmark complet + implications politiques).

---

2. MATERIALS & METHODS

2.1 Study area & sampling design

- Ajouter une **figure cartographique** de la zone Vina avec localisation dans le Cameroun/ASS
- Décrire le protocole d'échantillonnage (stratifié ? aléatoire ?)
- Justifier N=354 via une **analyse de puissance statistique a posteriori**

2.2 Data collection & variable description

- Tableau complet des 30 variables (nom, type, unité, statistiques descriptives)
- **SUPPRIMER "profit/ha" comme prédicteur** ou le reconstruire à partir de coûts de production (pas du rendement × prix) pour éviter la circularité. Expliquer explicitement comment le profit est calculé.

2.3 Data preprocessing (condensé)

- Résumer les étapes en 1 paragraphe + flowchart (Figure 1)
- Mentionner l'**analyse de sensibilité aux valeurs manquantes** (% par variable)

2.4 ML models (restructuré)

- **Supprimer les descriptions algorithmiques trop longues** (equations déjà publiées). Remplacer par un tableau de synthèse : modèle, hyperparamètres, justification du choix.
- **Ajouter SHAP (SHapley Additive exPlanations)** pour l'interprétabilité — c'est la norme 2024-2025 en ML agricole.

2.5 Validation framework (renforcé — point crucial)

- **Schéma de validation en 3 niveaux** :
  - Niveau 1 : Train/test split 80/20
  - Niveau 2 : 5-fold CV sur le train set
  - Niveau 3 : **Simulation de bruit** (Monte Carlo, N=1000 perturbations gaussiennes) pour tester la robustesse aux erreurs de mesure non calibrée

2.6 Performance metrics (corriger l'erreur fatale)

- **Supprimer TOUS les métriques de classification** (Recall, Precision, F1, Specificity)
- Utiliser UNIQUEMENT : R^2, RMSE, MAE, MAPE, NSE (Nash-Sutcliffe Efficiency)
- Ajouter le **Diebold-Mariano test** pour tester si les différences de performance sont statistiquement significatives

---

3. RESULTS (restructuré en 4 blocs logiques)

3.1 Descriptive statistics & variable importance

- Tableau 1 : Stats descriptives des variables (mean, SD, min, max, skewness)
- Figure 1 : Carte de la zone d'étude
- Figure 2 : SHAP beeswarm plot (pas juste le bar chart d'importance)
- **Adresser explicitement la circularité profit/ha** : montrer que le modèle reste performant même sans cette variable (ablation study)

3.2 Full-sample model performance

- Tableau 2 : Benchmark complet (tous les modèles, toutes les métriques)
- Figure 3 : Scatter plots observed vs predicted (conserver)
- **Ajouter Diebold-Mariano test en tableau** pour confirmer que XGBoost > RF de façon statistiquement significative

3.3 Cross-validation & generalizability

- Tableau 3 : CV results (corriger ANN manquant)
- Figure 4 : Violin plots des distributions de R^2 en CV (plus riche que les moyennes seules)
- Figure 5 : Learning curves (conserver + interpréter le convergence point)

3.4 Robustness to measurement error (SECTION NOUVELLE — killer feature)

- Résultats de la simulation Monte Carlo : comment se dégradent les R^2 de chaque modèle quand on ajoute du bruit gaussien σ = 5%, 10%, 20% aux rendements ?
- Message : XGBoost et RF sont les plus robustes → justifie leur recommandation dans des contextes à données imprécises.

3.5 Residual analysis

- Figure 6 : Boxplots des résidus (conserver)
- Figure 7 : Distribution des résidus XGBoost + Stacking (conserver)
- **Ajouter test de normalité des résidus** (Shapiro-Wilk) et d'hétéroscédasticité (Breusch-Pagan)

---

4. DISCUSSION (totalement restructurée — la section la plus importante pour un IF>5)

4.1 Socioeconomic data as a viable alternative to spectral/climate data

C'est l'argument le plus fort de votre article. Développer en confrontant à la littérature internationale. Pourquoi ça marche ? Quelles conditions sont nécessaires ?

4.2 Why do ensemble methods dominate ? (mécanismes, pas juste constats)

Expliquer via la théorie bias-variance. XGBoost capture les interactions non-linéaires profit×fertilisation×superficie que les modèles linéaires ratent.

4.3 The profit circularity debate (aborder frontalement)

Admettre le risque, montrer les résultats de l'ablation study (section 3.1), conclure sur la robustesse.

4.4 Scalability & policy implications (section nouvelle)

Si N=354 dans un district donne R^2=0.93 en CV, qu'est-ce que cela implique pour :

- Les ministères qui collectent des données d'enquête chaque année ?
- Le déploiement d'outils d'aide à la décision low-cost au Cameroun / en ASS ?
- La conception des prochaines enquêtes agricoles (quelles variables collecter en priorité) ?

4.5 Limitations & future research

Être **proactif** : N limité, données statiques, absence validation externe → annoncer le plan (étude longitudinale, validation multi-zone, intégration données météo comme robustness check).

---

5. CONCLUSION (3 paragraphes, pas de liste)

P1 — Réponse directe aux 3 questions de recherche **P2** — Implication scientifique (paradigme socio-économique vs spectral) **P3** — Implication pratique + appel à l'action pour les décideurs africains

---

📋 **CHECKLIST DES 10 ACTIONS PRIORITAIRES**

#	Action	Impact	Difficulté
1	Supprimer tous les métriques de classification → métriques de régression uniquement	🔴 Fatal si non corrigé	Facile
2	Ajouter ANN dans le tableau CV	🔴 Fatal si absent	Moyenne
3	Adresser la circularité profit/ha (ablation study)	🔴 Bloquant	Moyenne
4	Repositionner le titre et l'abstract sur la contribution socio-économique	🟠 Très fort	Facile

#	Action	Impact	Difficulté
5	Ajouter SHAP values pour l'interprétabilité	🔴 Attendu en 2025	Moyenne
6	Ajouter simulation Monte Carlo (robustesse bruit de mesure)	🔴 Différenciateur	Difficile
7	Diebold-Mariano test pour significativité statistique des comparaisons	🟡 Rigoureux	Moyenne
8	Carte de la zone d'étude	🟡 Standard	Facile
9	Tableau descriptif complet des 30 variables	🟡 Standard	Facile
10	Réécrire la Discussion avec les implications politiques	🟡 Valeur ajoutée	Facile

---

🎯 **REVUES CIBLES RECOMMANDÉES (par ordre de pertinence)**

Revue	IF	Justification
Computers & Electronics in Agriculture (Elsevier)	~8.3	Parfaite adéquation thématique, publie beaucoup de benchmarks ML avec données terrain
Agricultural Systems (Elsevier)	~6.6	Valorise les implications politiques et les contextes ASS
Field Crops Research (Elsevier)	~5.8	Si vous ajoutez l'angle agronomique
Frontiers in Sustainable Food Systems	~4.7	Open access, bon pour l'Afrique, moins de barrières
PLOS ONE	~3.7	Fallback solide si les autres rejettent

En résumé : Votre article contient une **idée vraiment forte** — démontrer que les données socio-économiques d'enquête suffisent pour prédire les rendements en ASS. Mais cette idée est noyée dans un cadrage générique "ML comparison paper". La transformation vers un IF>5 nécessite 3 semaines de travail : corriger les erreurs de fond (métriques classification, ANN manquant), ajouter l'analyse SHAP et la simulation Monte Carlo, et réécrire l'Introduction + Discussion autour du paradigme socio-économique. C'est faisable, et le résultat peut être vraiment impactant pour la communauté agricole africaine.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Dataofmaizefarmers.xlsx
- SURVEYSHEET.pdf