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.

Predicting Sorghum Yield in Data-Scarce and Conflict-Affected South Sudan Using Machine Learning Techniques.

John Karongo

johnkarongo@hotmail.com

International Committee of the Red Cross https://orcid.org/0009-0009-8942-5147

Joseph Mwaniki

School of Mathematics - University of Nairobi

John Ndiritu

School of Mathematics - University of Nairobi

Victor Mokaya

CSV Research Ltd https://orcid.org/0009-0001-0056-9921

Biological Sciences - Article

Keywords: Self-declared, sorghum yield, conflict, machine learning, Sorghum bicolor, L. data scarcity, Remote sensing

Posted Date: August 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4837792/v1

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

Additional Declarations: There is **NO** Competing Interest.

Predicting Sorghum Yield in Data-Scarce and conflict-Affected South Sudan Using Machine Learning Techniques

John Karongo[1,2*], Joseph Ivivi Mwaniki[2], John Ndiritu[2], Victor Mokaya[3]

[1] International Committee of the Red Cross, ICRC. Regional Delegation, Rue J47, BP, 459, Abidjan, Ivory Coast.
[2] School of Mathematics, University of Nairobi, Box, 30197, 00100, Nairobi, Kenya.
[3] CSV Research ltd, 643, Nairobi, 90131, Nairobi, Kenya.

*Corresponding author(s). E-mail(s): johnkarongo@hotmail.com;
Contributing authors: jimwaniki@uonbi.ac.ke; jndiritu@uonbi.ac.ke;
vmandela99@gmail.com;

Abstract

South Sudan continues to experience conflict. Agriculture production remains a crucial food and economic security feature for the overwhelming majority of the population and Sorghum is a key staple crop. In this country, sorghum yield estimations and predictions are of great interest to support interventions from the humanitarian sector, government policies, and for small-scale farmers' food and economic security. Ongoing conflict, limited access, and poor infrastructure suggest remote sensing and modeling as credible alternative for sorghum yield prediction in this country.

This research compares five popular regression techniques, namely Random Forest(RF), Decision Tree, Extreme Gradient Boosting (XGBoost), Support Vector Machine Learning(SVM) and Artificial Neural Network(ANN) Regressor for predicting sorghum crop yield in conflict setting in Upper Nile and Werstern Barh Al Gazal. The study uses 4 years soghum yield data, remotely sensed weather patterns, and above ground biomass proxies including NDVI, EVI, LAI characteristics as inputs for model training and evaluation. Preprocessing method like Fillna is used to handle missing data. The performance of each model is evaluated using metrics like RMSE, MSE, MAE, and $\boldsymbol{R^2}$. This study wanted to assess the

influence of conflict on small-scale farmers' sorghum yield prediction using remote sensing, climate data, and vegetation proxies in a data-scarce environment.
No research so far was done to investigate the impact of conflict on yield prediction in South Sudan. We constructed a modeling framework designed to incorporate data on the likelihood of conflict occurrences sourced from both Uppsala University and Acled, farmers perception on conflict, soil data and remotely sensed vegetation proxies to predict sorghum yield. Machine learning models were used to predict end-of-season sorghum yield and results show the Random Forest model yielded best combination of metrics with an RMSE of 176.27 Kg/ha and an $\boldsymbol{R^2}$ of 58 percent, confirming competitive performance. Outperforming other models, XGBoost records the best RMSE of 171.68 Kg/ha and an $\boldsymbol{R^2}$ of 59 percent. Renowned for its efficacy, XGBoost exceled in capturing intricate relationships within self-declared sorghum yield data, indicating its potential for accurate predictions. Both SVM and ANN achieved relatively lower performance metrics, with an RMSE of 188.95 Kg/ha and an $\boldsymbol{R^2}$ of 50.6 percent for SVM and for ANN a RMSE of 225kg/ha and an $\boldsymbol{R^2}$ of 52.5 percent. These findings for Decision trees, Random Forest, and XGBoost are significant in predicting sorghum yield in South Sudan. The variable importance reveals conflict probability was not a significant factor and did not influence sorghum yield prediction. Cultivated land size was the most significant predictor of sorghum yield in the context of South Sudan. These findings provide insights into the potential efficacy and limitations of machine learning for predicting sorghum yield in conflict settings in support to agricultural planning and humanitarian relief interventions.

Keywords: Self-declared sorghum yield conflict machine learning Sorghum bicolor L. data scarcity Remote sensing

1 Introduction

In South Sudan, the cultivation of sorghum plays a pivotal role in addressing food insecurity. But, Increasing population and conflict have imposed challenges in the agriculture sector in terms of productivity and food security.

The accurate estimation of sorghum production in South Sudan emerges as a critical imperative Dorosh et al. [1], Sorghum being an important staple crop in this country and the region.

With challenges related to instability and decades of insecurity and field access, remote sensing and modeling represents an opportunity to turn data into insights that will enable humanitarian actors and decision-makers to make strategic plans to meet people needs and ensure people's food security. Humanitarian actors, government, and farmers in South Sudan will benefit from accurate yield estimates to make strategic decisions that impact both economic outcomes, food security and effective humanitarian aid. In this country, the impact of conflicts on agricultural productivity cannot be understated. The interplay between conflict probability and crop yield adds a layer of intricacy to decision-making processes [2], which calls for a comprehensive approach that takes into account both the immediate challenges and ensures long-term sustainability. In this context of extensive farming systems with low inputs

low outputs, climate factors emerge as crucial determinants of sorghum yield at small-scale farmers. South Sudan's climatic conditions, characterized by high temperature variations, high variations in rainfall patterns, soil moisture, evapo-transpiration exert direct influences on crop growth and development. Understanding and quantifying the influence of these climatic variables are essential for unraveling the intricate relationship between weather patterns and sorghum production. The use of traditional approaches for crop yield estimations is a time-consuming process and very risky to be conducted within volatile and unpredictable circumstances due to chronic instability and security that hinder the execution of large-scale surveys, particularly in South Sudan. In this country, existing approaches for crop yield estimation mostly rely on small scale production surveys. The need for sorghum yield prediction especially when conventional large-scale surveys can not be conducted due to conflict and lack of access, existing limited data is of great importance [3, 4].

In the era of digital agriculture [5], characterized by smart farming tools and technologies Wolfert et al. [6], Balakrishnan and Muthukumarasamy [7], the integration of advanced methodologies becomes a transformative avenue for addressing these challenges. Artificial intelligence, big data analytics, cloud computing, machine learning (ML), [8], and sensor technologies collectively provide a comprehensive platform for data collection, storage, and analysis [9]. Unfortunately, these technologies are far from reaching full penetration in terms of capacity building in South Sudan. Machine learning involves the development and examination of algorithms capable of learning from data. These algorithms function by constructing models from input data and utilizing them for making predictions or decisions, rather than relying solely on explicitly programmed instructions. With minimal human intervention, machine learning provides enhanced decision-making support [10, 11]. In this research work the machine learning algorithms used are Artificial Neural Networks(ANN), Support Vector Machines (SVM), Random Forest(RF), Extreme Gradient Boost(XGBOOST), and Decision trees(DT), to predict sorghum yield in conflict setting. These ML models stand out for their ability to navigate the complexities of both linear and non-linear agricultural data Liakos et al. [12], making them invaluable for sorghum yield predictions in the South Sudanese context[11, 13].

However, the efficacy of ML predictions is contingent upon the quality of the input data, model representation, and the specific target variable—in this case, sorghum yield. In this study, an incremental technique is employed, commencing with soil data and progressively incorporating weather data, conflict probability, and other remotely sensed vegetation proxies factors. The aim is to assess the impact on sorghum yield predictions across a spectrum of ML algorithms, providing a nuanced understanding of the interplay between conflict, climate, and sorghum production in South Sudan.

In summary, this research endeavors to contribute valuable insights into optimizing the prediction accuracy of sorghum yield in South Sudan. By addressing the multi-faceted challenges posed by conflict, climate factors, and the intricate dynamics of crop production, the study aspires to equip agricultural stakeholders, humanitarian actors, and government with the tools they need to make informed decisions, ensuring people's food needs are met in the country. We adopted five machine-learning algorithms for predicting sorghum yield, including artificial neural networks(ANN), Support Vector

Machines (SVM) [14, 15], Random Forest(RF), Extreme Gradient Boost(XGBOOST), and Decision trees(DT). Our primary goals included: (1) establishing a framework for predicting sorghum yield; (2) identifying the most effective machine learning algorithms for sorghum yield prediction; and (3) investigating the variations in sorghum yield prediction along with the relative significance of variables[16, 17].

The objective of this research is to suggest an integrated approach for predicting sorghum crop yield at the scale of small-holder farmers, considering the uncertainty in complex and diverse farming systems impacted by persistent conflict. The proposed method involves utilizing remotely sensed data, identifying optimal crop-specific yield estimators from a couple of vegetation status proxies over time, and incorporating a stochastic combination of these outcomes along with their associated uncertainties in a context of protracted conflict.

2 Materials and Methods

2.1 Study Area

2.1.1 Upper Nile State

Upper Nile is one of the states in South Sudan and is positioned at coordinates 09°46'N 32°44'E, with its headquarters located in Malakal. The state comprises 13 counties and is home to an estimated population of 964,353 inhabitants, as per the South Sudan Census of 2008(https://unstats.un.org/unsd/demographic/sources/census). Spanning an area of 77,823.42 square kilometers, the region is primarily engaged in the cultivation of sorghum, wheat, millet, and sesame. These crops are grown through either rain-fed or irrigated methods. The cultivation period in this state goes from April to November, encompassing land preparation in April/May, planting in July or early August, and harvesting by November at the latest.The climate there is characterized by a unimodal rainfall season, with annual rainfall ranging from 500 mm to 650 mm with the peaks in July and August It's worth noting that due to security concerns, access to many parts of Upper Nile state is problematic.

2.1.2 Western Bahr El Ghazal State

Western Bahr el Ghazal located at 07°53'N 25°52'E is another state with Wau as its capital city. It has an area of 91,079 kilometers squared and is the least populous state (333,431 persons) in South Sudan, according to the census conducted in 2008. The majority of this state is situated along the northern section of the Ironstone plateau, featuring red laterite soils with occasional alluvial deposits in the low-lying areas.
The region typically receives ample rainfall, providing support for more than 180 growing days conducive to cereal production. In general, crop planting is completed by mid-May when the soil moisture is good enough after 2-3 good rains to initiate the germination. Sorghum usually matures after 5 months and harvest happens in late November beginning of December. There is a general dependence on carryover seeds. Western Bahr El Gazal and Upper Nile states are two of the 11 states that form the Republic of South Sudan. In these two states, diversified crops are cultivated but Sorghum (Sorghum bicolor) is the predominant crop.



Fig. 1 Map of South Sudan showing Upper Nile and Western Bahr El Ghazal States. Source: Google

2.2 Data Sources

2.2.1 Remote sensing data

We collected three types of vegetation index data (VI), the normalized difference vegetation index (NDVI), the Enhanced Vegetation Index (EVI) and the Leaf Area Index (LAI) from MODIS product. A vegetation index can monitor the dynamic change of vegetation Motohka et al. [18]. Currently, many studies have shown that NDVI and EVI have a good correlation with crop yield [16, 19] and the combination of NDVI and EVI can provide more crop information, which can efficiently use for crop yield prediction [20].In this study, we used six years (2015-2020) of remote sensing data from MODIS products.

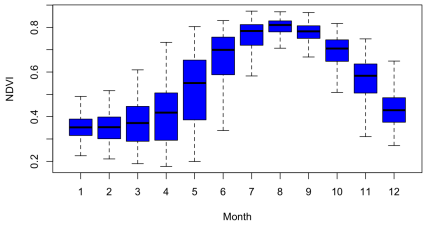


Fig. 2 NDVI distribution Barh El Gazal

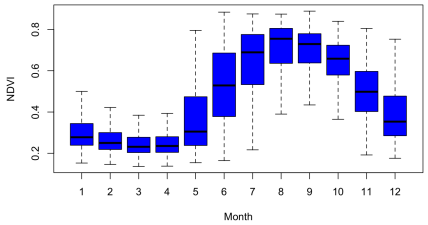


Fig. 3 NDVI distribution Upper Nile

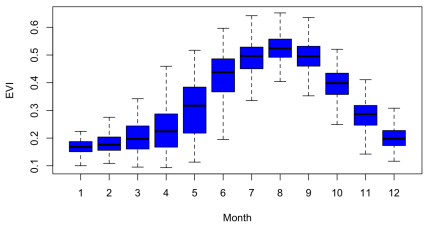


Fig. 4 EVI distribution Barh El Gazal

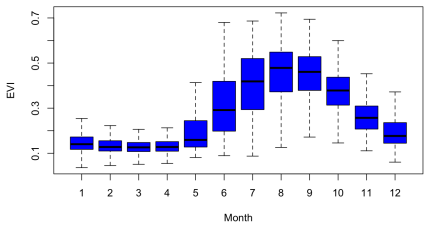


Fig. 5 EVI distribution Upper Nile

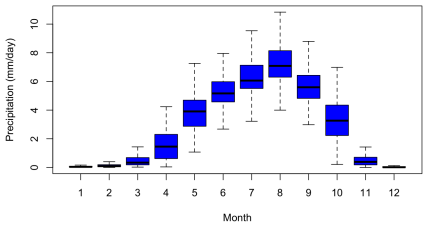


Fig. 6 Precipitation distribution Barh El Gazal

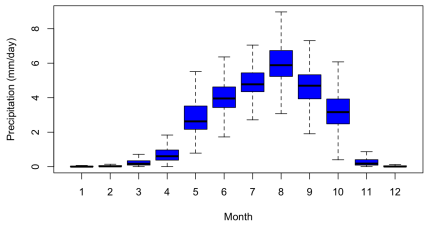


Fig. 7 Precipitation distribution Upper Nile

2.2.2 Climate and soil Data

Climate variables including Precipitation, Daytime land temperature, and Evapotranspiration are important drivers for crop production. Droughts and Extreme temperatures have adverse impacts on crops in the frame of global climate

change. In this study, we used daily temperature, precipitation and evapotranspiration data from EARTH Data (Science for a Changing World USGS - NASA) to predict sorghum yield [21, 22]. Soil moisture data were also collected https://lpdaac.usgs.gov/products/mod11a1v061/

Soil is a crucial factor affecting crop yield. Soil moisture data were collected from the Earthdata NASA, the National Snow and Ice Data Center(NSIDC) anchor-data-access-tools link

2.2.3 Sorghum yield Data

Field surveys were conducted during 4 years and self-declared post harvest sorghum yield data (2018-2021) were collected in 15 villages of 2 states including 8 villages in Upper Nile and 7 Western Bahr El Ghazal states. A total of 2,100 small-scale farmers were randomly selected and interviewed through a network of surveyors from within local communities.

2.2.4 Conflict Data

Conflict data were used from Uppsala University conflict database https://ucdp.uu.se. Uppsala is the world leading provider of data on armed conflict and organized violence. This conflict data was confronted against ACLED(Armed Conflict Location and Event Data - www.acleddata.com) ACLED is an NGO specialized in disaggrated data collection, analysis and crisis mapping. In addition, farmers' perception of conflict and its impact on their cultivated land and sorghum production data were collected.

2.3 Machine learning methods for estimating sorghum yield

2.3.1 Decision trees

A decision tree machine learning model possesses a structure resembling a tree, and it predicts probable outcomes [23, 24], assesses resource costs, evaluates utility costs, and considers potential implications. [25, 26]. They operate by recursively partitioning the input space based on feature values, forming a tree structure. It involves the following steps:

- *Node Splitting:* Commencing with the complete dataset, the process involves selecting the feature that most effectively segregates the data according to a specific criterion, such as Gini impurity or information gain.
- *Data Partitioning:* The next step involves partitioning the data into subsets based on the selected feature.
- *Recursive Process:* Iterate through the process for each subset until a predefined stopping criterion is fulfilled, such as reaching the maximum depth or attaining a minimum number of samples per leaf.

The ultimate prediction is derived by averaging or aggregating the predictions from individual trees, typically through a majority vote mechanism.

2.3.2 Random Forest algorithm (RFs)

Random Forests, which is also known for classification and regression, is a non-parametric supervised learning ensemble method[23, 27]. It accomplishes these tasks by constructing predictive models using multiple learning algorithms. To ensure the forests are uncorrelated, the algorithm employs randomization in creating each tree. These trees, in turn, contribute to the forest's predictive capabilities, enhancing the accuracy of decision-making Senagi et al. [28]. To foster uncorrelated forests, the algorithm utilizes randomization in the construction of each tree. Subsequently, these trees employ the predictive powers of the forest to make accurate decisions. [29, 30].

2.3.3 Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful gradient-boosting algorithm known for its efficiency and high performance. It builds a series of weak learners sequentially and corrects errors made by previous models. It involves these major steps:-

- *Sequential Training:* Train a weak learner on the dataset.
- *Residual Computation:* Compute the residuals between the predicted and true values.
- *Iterative Process:* Fit a new weak learner to the residuals, and repeat until a predefined number of trees is reached.

The prediction from the ensemble is the sum of predictions from all individual trees, weighted by a learning rate Ren et al. [31]:

$$F(x) = \sum_{i=1}^{N} \eta f_i(x)$$

2.3.4 Support Vector Machines (SVM)

SVM is a powerful model for both classification and regression tasks. It finds the hyperplane that maximally separates data points of different classes in the feature space [32, 33]. It involves these steps:-

- *Margin Maximization:* Find the hyperplane that maximizes the margin between classes.
- *Kernel Trick:* Introduce a kernel function for non-linear separation.
- *Classification:* Classify new data points based on their position relative to the hyperplane.

The decision function for SVM can be expressed as Das et al. [34]:

$$f(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

2.3.5 Artificial Neural Networks(ANN)

ANN is a computational model which mimics the human nervous system. ANN is commonly applied to predict crop yield (Akbar et al. [35]. They consist of layers of

interconnected nodes (neurons) that learn hierarchical representations of data Yildirim et al. [36]. Neural network has become a central concept in modern machine learning.

A network of neurons can be constructed by linking multiple neurons together in the sense that the output of one neuron forms an input to another. A simple model for such a network is the *multilayer perceptron* as introduced by Torkashvand et al. [37]

A neuron-like processing unit can be written as follow:

$$\alpha = \phi\left(\sum_j w_j x_j + b\right)$$

Where the x_j are the inputs to the unit, the w_j are the weights, b is the bias, ϕ is the nonlinear activation function, and a is the unit's activation.

- Units with linear regression uses a linear model, so $\phi(z) = z$.
- *Feedforward:* Pass input through layers to generate an output.
- *Backpropagation:* Compute error, and update weights using gradient descent.
- *Activation Functions:* Activation functions introduce non-linearity to the model Khaki et al. [38].

In this research, the input layer consists of several neurons, including remotely sensed data, climate and soil moisture data, and conflict data. [39, 40].

2.4 Model Evaluation

Assessment of prediction accuracy of the models on both training and testing sets were based on three common statistical metrics [41]: coefficient of determination (2.4) R^2, root mean square error RMSE, (1), and the mean absolute error MAE (2).

R^2 is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables. RMSE and MAE, on the other hand, quantify the average errors between predicted and actual values. [41]. The average values of independent metrics across all models were also analyzed

$$R^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- x_i and y_i are the individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y data points, respectively.
- n is the number of data points.

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

where: n is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{2}$$

where: n is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

3 Results and Discussion

3.1 Descriptive Statistics

In this section, we present a table and a figure, as shown in figure 9 and Table 1, the variables under consideration are land size, sorghum yield, evapotranspiration, soil moisture, and precipitation, each of which plays a pivotal role in sorghum cropping.

	Land size	sorghum yield	Evapotranspiration	soil moisture	Precipitation
Mean	0.633	366.03	783.17	0.2822	2.517
std	0.461	292.29	4949.49	0.1027	2.463
min	0.1	0	0	0.073	0
25%	0.4	150	8	0.2021	0.088
50%	0.5	300	16.68	0.2674	2.071
75%	0.8	450	26.3	0.3529	4.57
Max	3	2345	32766	0.504	10.706

Table 1 Summary statistics of variables

A broad range of cultivated land and sorghum yield was observed in the two states. The maximum cultivated land size was 3ha, indicating the presence of relatively larger land areas in the Upper Nile state. The mean cultivated land size in both states was approximately 0.633 ha, The standard deviation for land size was at 0.461ha, suggesting a moderate degree of variability in cultivated land sizes across the two regions. While the sorghum yield averaged around 366.03 kg/ha with a maximum yield at 2'345kg/ha, showcasing the potential for high-yield outcomes.The standard deviation for sorghum yield was 292.29 kg/ha, indicating a considerable range of yield outcomes. Precipitation data reveals a maximum value of 10.706 mm/day reflecting the highest observed value, and an average of 2.517 mm/day. The precipitation standard deviation was of 2.463 mm/day reflecting the variability in regional precipitation patterns. Together, these descriptive statistics offer a comprehensive overview of the agricultural and environmental characteristics of the two states.

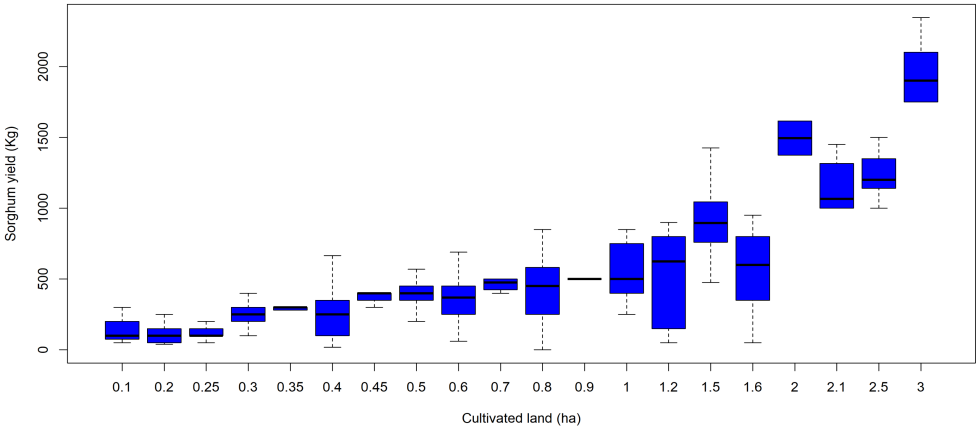


Fig. 8 Sorghum Yield against cultivated land in hectare

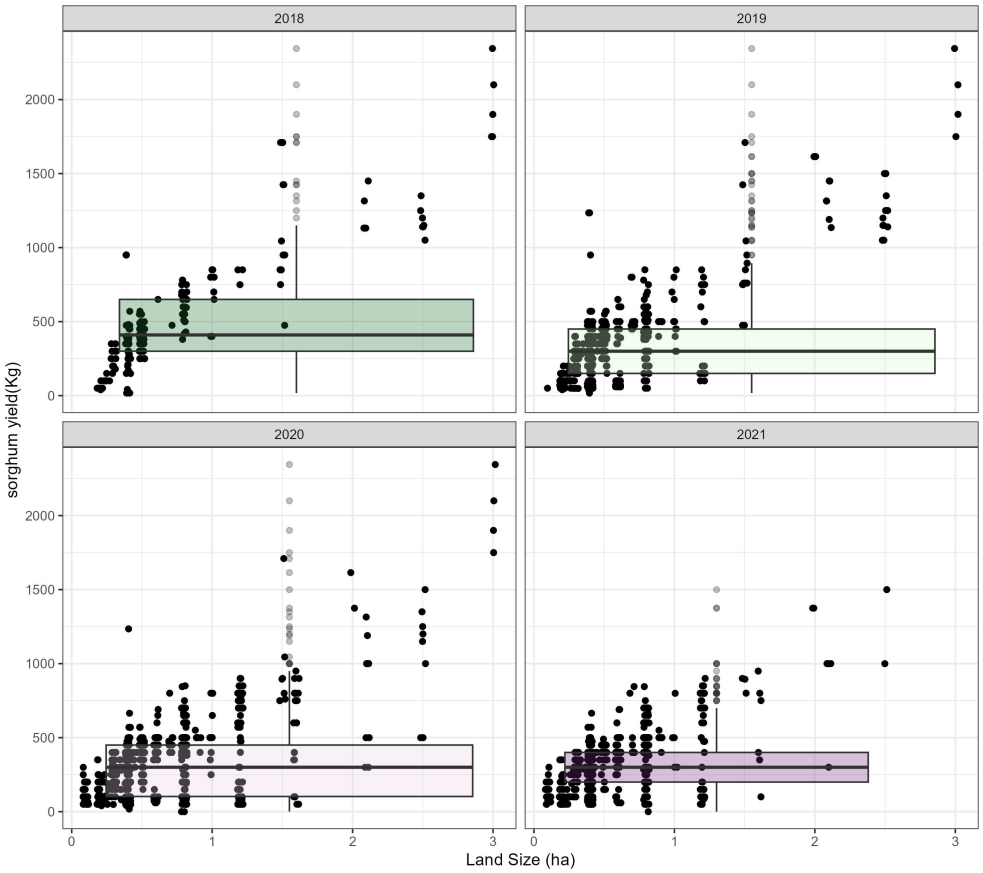


Fig. 9 Descriptive Statistics of yield according to the land size and across years

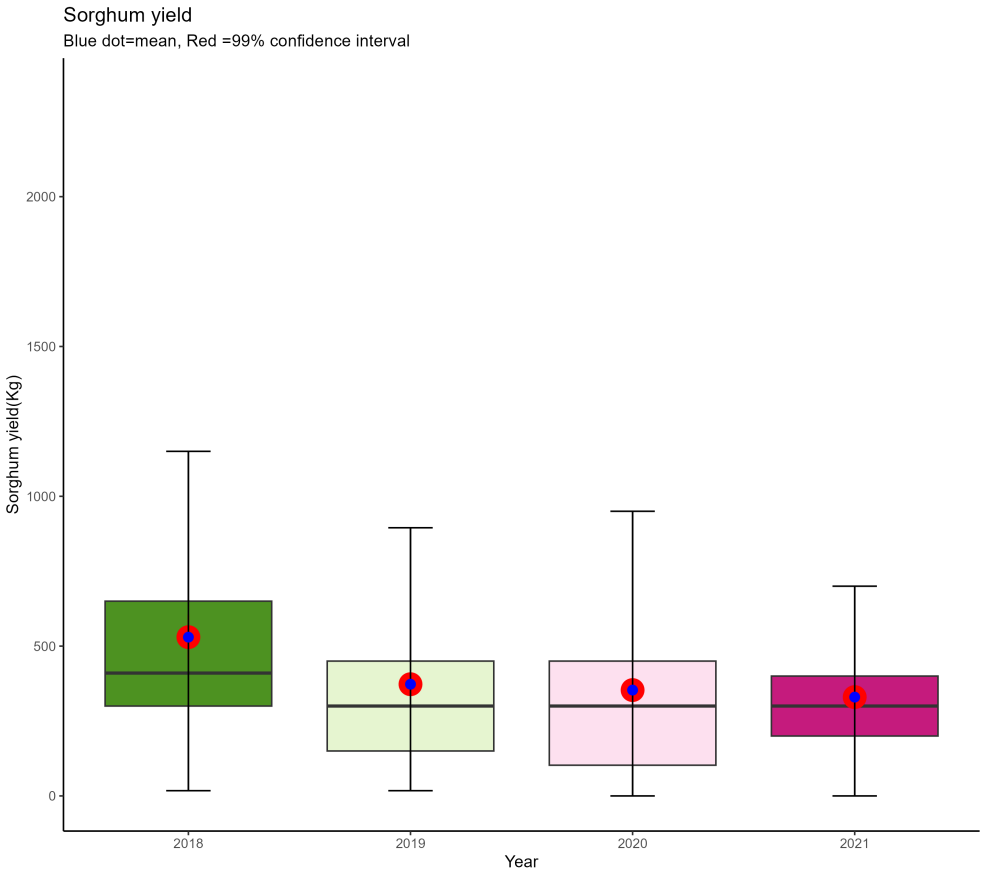


Fig. 10 Sorghum yield in Kg across years

3.2 Correlation

The correlation table 2 and figure 11, illuminate the intricate relationships among key parameters in the studied region, offering insights into the direction, strength, and statistical significance of these connections.

A strong positive correlation of 0.75 is observed between land size and sorghum yield, indicating a substantial and statistically significant association. The larger cultivated land areas tend to be the higher the sorghum yield is confirming the practice in low-input-low-output farming system.

The correlation between sorghum yield and precipitation is moderate but lacks statistical significance ($p = 0.103$). Soil moisture and precipitation exhibit a weak positive correlation of 0.05, with a p-value of 0.297.

Parameter1	Parameter2	r	95% CI	t(2094)	p
Landsize	SorgYield	0.75	[0.73, 0.77]	51.90	<.001***
Landsize	Evapotranspiration.	-0.04	[-0.08, 0.00]	-1.92	0.384
Landsize	Soil_Moisture.	0.02	[-0.03, 0.06]	0.73	>.999
Landsize	Precipitation.	-4.87e-03	[-0.05, 0.04]	-0.22	>.999
SorgYield	Evapotranspiration.	-0.02	[-0.06, 0.02]	-0.97	>.999
SorgYield	Soil_Moisture.	8.01e-03	[-0.03, 0.05]	0.37	>.999
SorgYield	Precipitation.	-0.06	[-0.10, -0.01]	-2.53	0.103
Evapotranspiration.	Soil_Moisture.	0.03	[-0.02, 0.07]	1.17	>.999
Evapotranspiration.	Precipitation.	-0.02	[-0.06, 0.03]	-0.79	>.999
Soil_Moisture.	Precipitation.	0.05	[0.00, 0.09]	2.09	0.297

Table 2 Correlation relationship between variables

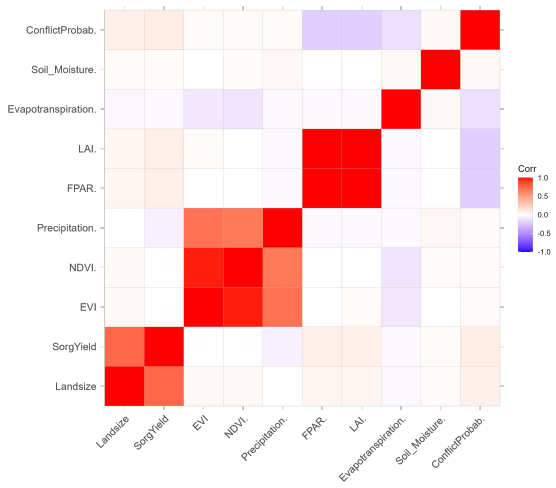


Fig. 11 Correlation of variables

3.3 Comparing training accuracy of sorghum yield on different models

In this study, five machine learning models were trained with the observed yields and variables of sorghum from 2018 to 2020. The evaluated results, based on the ten-fold cross-validation, were summarized according to different models and time windows in figure 12 and 13. Comprehensively considering three evaluation indicators (R^2, RMSE, and MAE), Decision trees, RF, XGBOOST and ANN models showed high accuracy, with all having high R^2 (> 0.7) indicating a significant relationship between the predicted and observed yied. Decision trees and XGboost both had an accuracy close to 0.8. Decision trees had the lowest RMSE (200.71 Kg/ha) while XGboost showed an RMSE of 252.75 Kg/ha. While having a good predictive accuracy (0.74) on the learning dataset, the ANN had the highest RMSE (304.68 Kg/ha) compare to all the models. SVM performed slightly poor with an accuracy of 0.68 and an RMSE of 234.92 Kg/ha. Thus, XGBOOST, RF, and Decision trees were more suitable for sorghum yield prediction than ANN and SVM. Moreover, we found that the training accuracy varied by time windows even with the same machine learning algorithm,

especially for the RMSE and MAE. However, the time windows showed less impact on R^2 values of XGBOOST, ANN, RF, and SVM. Finally, three algorithms (XGBOOST, RF, and Decision trees) were selected to establish prediction models for sorghum yield.

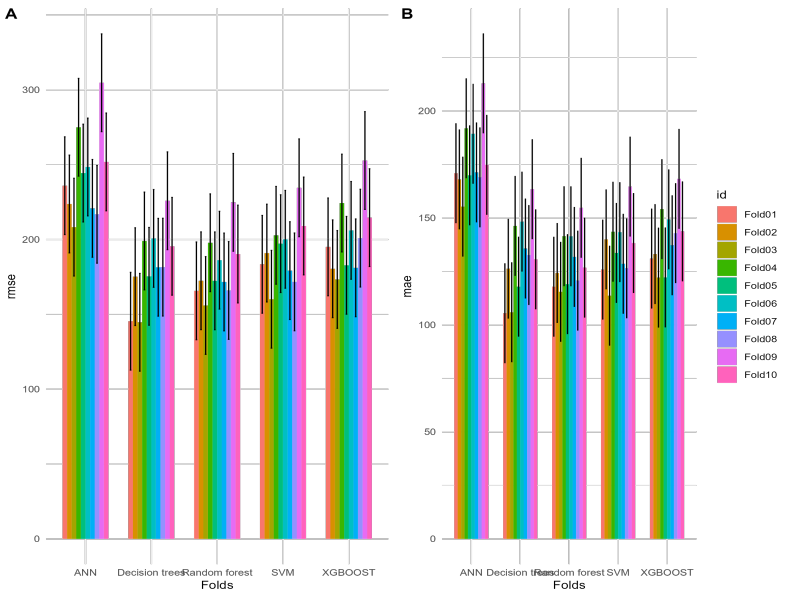


Fig. 12 RMSE(A), and MAE(B) metric scores of five models of sorghum yield based on ten-fold cross-validation results in different time windows. (The unit of RMSE and MAE is kg/ha)

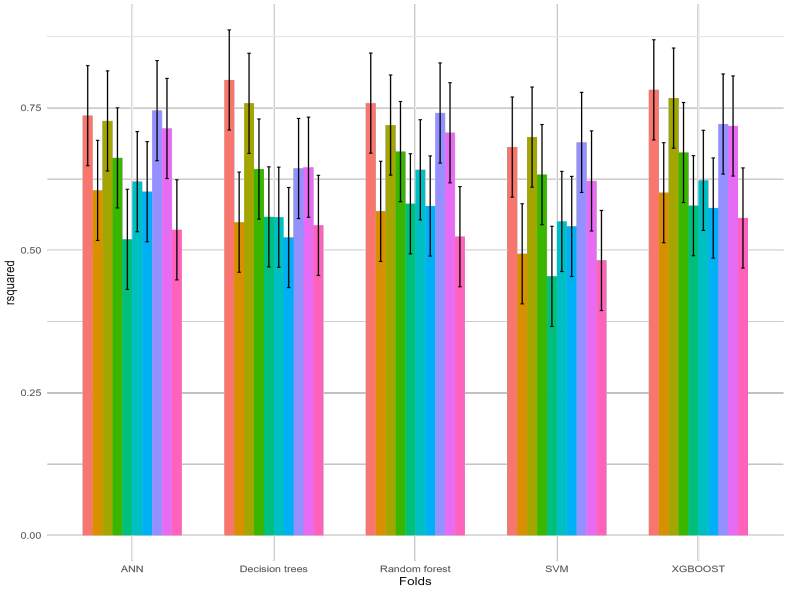


Fig. 13 R squared metric scores of five models of sorghum yield based on ten-fold cross-validation results in different time windows.

3.4 Sorghum yield predictions

Based on the trained models of DECISION TREES, ANN, RF, SVM, and XGBOOST in Section 3.3, sorghum yield in the 2 states in 2021 were predicted. The residuals of the prediction results of these models all passed the Kolmogorov-Smirnov test and obeyed normal distribution, which showed that these regression models were acceptable (figure 14). The scatter diagrams of the predicted and observed yields for the different models are shown in Figure. 14. We found that the predicted and observed yields showed a good linear fit with R^2 of about 0.5 for all the 5 models. Random Forest, Decision trees and XGboost showed higher accuracy in predicting sorghum yield over 57 percent. Such results indicated that the three machine learning models can predict sorghum yield with quite good accuracy in following order $XGBOOST > DECISIONTREE > RANDOMFOREST >$. Furthermore, the prediction areas were overestimated for low yields observed with smaller deviations, while underestimated for high observed yields with relatively greater deviations.

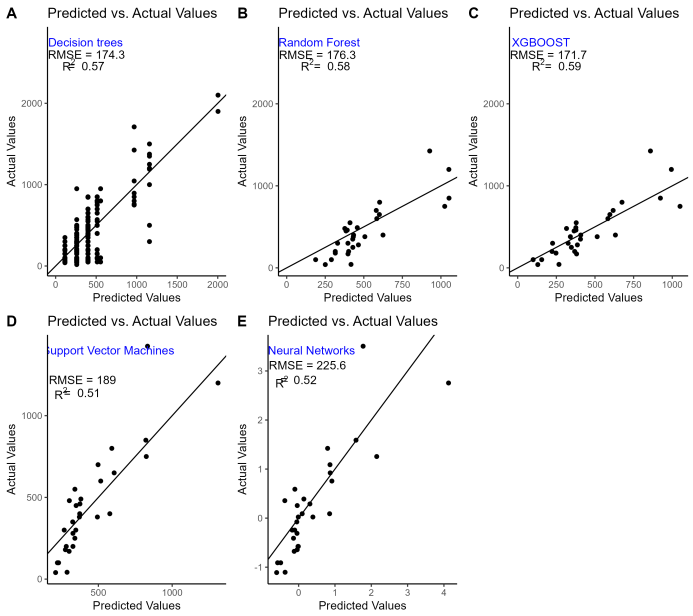


Fig. 14 Scatter plots of predicted yield vs observed sorghum yield

3.5 Impact of different models on prediction accuracy

Based on the RMSE and R^2 metrics, XGBoost emerged as the most effective model for predicting sorghum yield using remote sensing data in South Sudan, providing the best balance between prediction accuracy and explanatory power. Decision Trees and Random Forest also showed moderate performance but were outperformed by XGBoost. In contrast, SVM and Neural Networks did not perform as well, indicating

that they may require further optimization or may not be as suitable for this specific prediction task under scarcity of data.

In this section as shown in table 3 and figure 15, we the metric results for the predicting sorghum yield in 2021.

Model	Training metrics		Testing metrics	
	rmse	R^2	rmse	R^2
Decision trees	200.71	0.799	174.31	0.573
Random Forest	224.82	0.758	176.27	0.576
xgboost	252.75	0.782	171.68	0.586
SVM	234.92	0.689	188.95	0.506
NN	304.68	0.745	225.62	0.525

Table 3 Best performance metrics of sorghum models after parameter tuning

Outperforming other models, XGBOOST model yielded the highest combination of best metrics with an RMSE of 171.68kg/ha and an R^2 of 58.6 percent.Renowned for efficiency, XGBoost excels in capturing intricate relationships within the sorghum yield data, indicating its potential for accurate sorghum yield prediction. DECISION TREE model had the second best competitive performance with a combination of RMSE of 174.31 kg/ha and R^2 of 57.3 percent demonstrating effectiveness in sorghum yield predictions. RANDOM FOREST exhibits a RMSE of 176.27 kg/ha and an R^2 of 57.6 percent. These findings are consistent with those of [42] on Maize in the Sahel region. These results are significant in predicting sorghum yield in South Sudan.

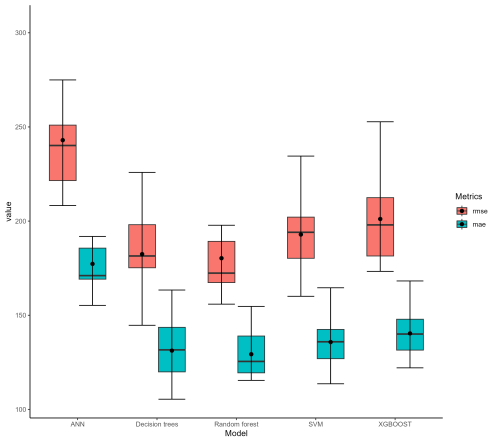


Fig. 15 Description of test metric for the models used for prediction. RMSE and MAE values

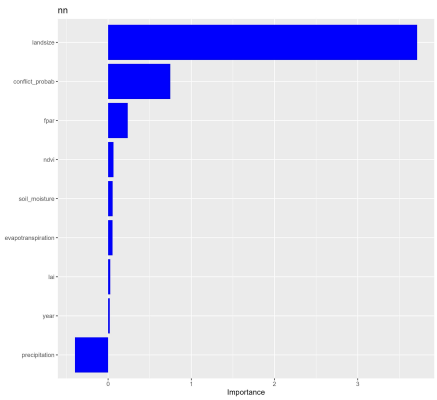


Fig. 16 Variable importance for ANN

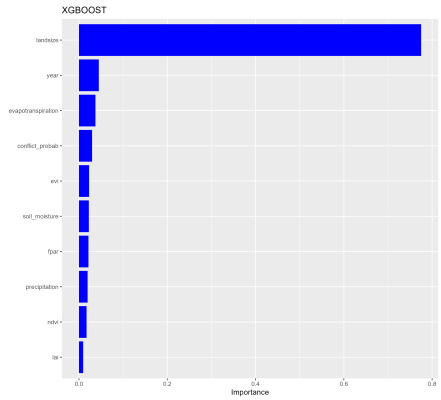


Fig. 17 Variable importance for XGBoost

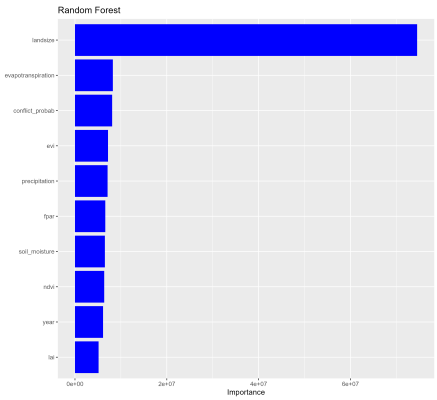


Fig. 18 Variable importance for Random Forest

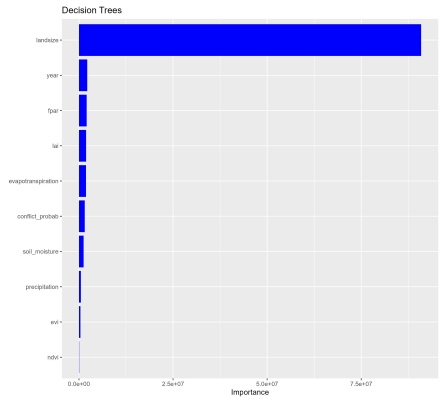


Fig. 19 Variable importance for Decision Trees

The variable importance of the decision tree, random forest, and XGboost are displayed in figure 16 and 18. All of the models show that land size and land moisture were the most important variables in predicting sorghum yield in South Sudan. Conflict probability does not appear to be a major factor in determining the sorghum yield.

The significance of a single factor cannot be evaluated in isolation from other factors. From this research XGBOOST, DECISION TREE and RANDOM FOREST machine learning offer a mean to gauge the relative importance of variables in predicting sorghum yield using vegetation proxies and soil moisture parameters [4, 43] in a conflict affected context such as South Sudan. Results underscore the vital role of vegetation proxies such NDVI, EVI and climate data in sorghum yield prediction,

aligning with prior research highlighting the significance of factors such as NDVI, precipitation, and temperature in sorghum yield prediction [19, 44]

This research found that land size is a pivotal variable and plays a substantial role in sorghum yield predictions in the context of South Sudan characterized by an extensive low-input-low-output sorghum farming system as shown in figure 18. Conflict probability, as an external factor, adds another layer of complexity to the predictive model but did not influence much sorghum yield prediction for all the models.

4 Conclusion

This study explored the potential for the estimation of sorghum yield using the 5 models (Decision trees, Random Forests, XGboost, SVM, and Neural Networks) in context of chronic conflict and data scarcity. The results of Sorghum yield modeling showed that DECISION TREES, XGBOOST and RANDOM FOREST predicted sorghum yield with good accuracy. Moreover, we investigated the prediction accuracy and found the models have respectively very good prediction accuracy of R^2 of 57 percent, 59 percent, 56.2 percent and lower RMSE($< 176kg/ha$)and MAE ($< 105kg/ha$). We also found that the prediction accuracy varied by land size, soil moisture, and algorithms, and that the regional difference also affect the sorghum yield prediction accuracy.

In conclusion, our investigation unveiled that the prediction accuracy was subject to variations based on factors such as land size, soil moisture levels, precipitations and the specific algorithms employed. Additionally, regional disparities in South Sudan emerged as a significant determinant influencing sorghum yield prediction accuracy. This research confirms land size, soil moisture and precipitations are key predictors of sorghum cropping in low-input-low-output farming system such as in South Sudan. These findings offer valuable insights for policymakers, agricultural practitioners, humanitarian actors, and researchers striving to enhance food security in South Sudan.

5 Acknowledgements

We would like to thank GIS colleagues and friends who contributed to download remotely sensed data and for the data treatment. The views expressed in this research paper are those of the authors and not the ICRC.

Declarations

- Funding
  This research received no external funding.
- Conflict of interest
  The authors declare no conflict of interest.
- Ethics approval and consent to participate
  Before interviews farmers were informed about the research and requested for their consent

- Consent for publication
  Yes
- Data availability
  Datasets used and/or analyzed during the current study will be made available from the corresponding author on reasonable request. Remotely sensed data are freely available online. Precipitation data, daytime land temperature and Evapotranspiration data, NDVI, LAI all available online.
- Author contribution
  Conceptualization, Design, methodology, John K. and Joseph I., John N., Investigation, John K., model development, validation, John K, Joseph I. and Victor M., writing—original draft, John K., Joseph I., Victor M., John N. All authors have read and agreed to the published version of the manuscript

References

[1] Dorosh, P.A., Rashid, S., Asselt, J.: Enhancing food security in south sudan: the role of markets and regional trade. Agricultural Economics **47**(6), 697–707 (2016)

[2] Vesco, P., Kovacic, M., Mistry, M., Croicu, M.: Climate variability, crop and conflict: Exploring the impacts of spatial concentration in agricultural production. Journal of Peace Research **58**(1), 98–113 (2021)

[3] Orr, A., Mwema, C., Gierend, A., Nedumaran, S.: Sorghum and millets in eastern and southern africa: facts, trends and outlook. ICRISAT (2016)

[4] Okeyo, S.O., Ndirangu, S.N., Isaboke, H.N., Njeru, L.K.: Determinants of sorghum productivity among small-scale farmers in siaya county, kenya. African Journal of Agricultural Research **16**(5), 722–731 (2020)

[5] Klerkx, L., Jakku, E., Labarthe, P.: A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS-Wageningen Journal of Life Sciences **90**, 100315 (2019)

[6] Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J.: Big data in smart farming–a review. Agricultural Systems **153**, 69–80 (2017)

[7] Balakrishnan, N., Muthukumarasamy, G.: Crop production-ensemble machine learning model for prediction. International Journal of Computer Science and Software Engineering **5**(7), 148 (2016)

[8] Ngo, V.M., Kechadi, M.-T.: Crop knowledge discovery based on agricultural big data integration. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing, pp. 46–50 (2020)

[9] Chergui, N., Kechadi, M.-T., McDonnell, M.: The impact of data analytics in digital agriculture: a review. In: 2020 International Multi-Conference On: "Organization of Knowledge and Advanced Technologies" (OCTA), pp. 1–13. IEEE, Berlin, Heidelberg (2020)

[10] Cedric, L.S., Adoni, W.Y.H., Aworka, R., Zoueu, J.T., Mutombo, F.K., Krichen, M., Kimpolo, C.L.M.: Crops yield prediction based on machine learning models: Case of west african countries. Smart Agricultural Technology **2**, 100049 (2022)

[11] Stas, M., Van Orshoven, J., Dong, Q., Heremans, S., Zhang, B.: A comparison of machine learning algorithms for regional wheat yield prediction using ndvi time series of spot-vgt. In: 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp. 1–5. IEEE, ??? (2016)

[12] Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: Machine learning in agriculture: A review. Sensors **18**(8), 2674 (2018)

[13] Crane-Droesch, A.: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters **13**, 114003 (2018)

[14] Gandhi, N., Armstrong, L.J., Petkar, O., Tripathy, A.K.: Rice crop yield prediction in india using support vector machines. In: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–5. IEEE, Berlin, Heidelberg (2016)

[15] Dang, C., Liu, Y., Yue, H., Qian, J., Zhu, R.: Autumn crop yield prediction using data-driven approaches: support vector machines, random forest, and deep neural network methods. Canadian Journal of Remote Sensing **47**(2), 162–181 (2021)

[16] Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y.: Crop yield forecasting on the canadian prairies using modis ndvi data. Agricultural and Forest Meteorology **151**(3), 385–393 (2011)

[17] Bansal, Y., Lillis, D., Kechadi, T.: Winter wheat crop yield prediction on multiple heterogeneous datasets using machine learning. In: 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 206–212. IEEE, Berlin, Heidelberg (2022)

[18] Motohka, T., Nasahara, K.N., Oguma, H., Tsuchida, S.: Applicability of green-red vegetation index for remote sensing of vegetation phenology. Remote Sensing **2**(10), 2369–2387 (2010)

[19] Huang, J., Wang, H., Dai, Q., Han, D.: Analysis of ndvi data for crop identification and yield estimation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(11), 4374–4384 (2014)

[20] Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D.: Machine learning in agriculture: A comprehensive updated review. Sensors **21**(11), 3758 (2021)

[21] Katimbo, A., Rudnick, D.R., Zhang, J., Ge, Y., DeJonge, K.C., Franz, T.E., Shi, Y.e.a.: Evaluation of artificial intelligence algorithms with sensor data assimilation in estimating crop evapotranspiration and crop water stress index for irrigation water management. Smart Agricultural Technology **4**, 100176 (2023)

[22] Mirzania, E., Kashani, M.H., Golmohammadi, G., Ibrahim, O.R., Saroughi, M.: Hybrid coot-ann: a novel optimization algorithm for prediction of daily crop reference evapotranspiration in australia. Theoretical and Applied Climatology **154**(1-2), 201–218 (2023)

[23] Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer, Berlin, Heidelberg (2000)

[24] Song, Y.Y., Ying, L.U.: Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry **27**(2), 130 (2015)

[25] Gavankar, S.S., Sawarkar, S.D.: Eager decision tree. In: 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 837–840. IEEE, Berlin, Heidelberg (2017)

[26] Burhan, H.A.: Crop yield prediction by integrating meteorological and pesticides use data with machine learning methods: An application for major crops in turkey. Ekonomi Politika ve Finans Araştırmaları Dergisi **7**(Özel Sayı), 1–18 (2022)

[27] Bali, N., Singla, A.: Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. Archives of Computational Methods in Engineering **29**(1), 95–112 (2022)

[28] Senagi, K., Jouandeau, N., Kamoni, P.: Using parallel random forest classifier in predicting land suitability for crop production. Journal of Agricultural Informatics **8**(3) (2017)

[29] Han, Q., Gui, C., Xu, J., Lacidogna, G.: A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm. Construction and Building Materials **226**, 734–742 (2019)

[30] Zhang, H., Zimmerman, J., Nettleton, D., Nordman, D.J.: Random forest prediction intervals. The American Statistician (2019). Online ahead of print

[31] Ren, Y., Ling, F., Wang, Y.: Research on provincial-level soil moisture prediction based on extreme gradient boosting model. Agriculture **13**(5), 927 (2023)

[32] Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J.: Prediction of winter wheat yield based on multi-source data and machine learning in china. Remote Sensing **12**(2), 236 (2020)

[33] Cubillas, J.J., Ramos, M.I., Jurado, J.M., Feito, F.R.: A machine learning model for early prediction of crop yield, nested in a web application in the cloud: a case study in an olive grove in southern spain. Agriculture **12**(9), 1345 (2022)

[34] Das, P., Jha, G.K., Lama, A., Parsad, R.: Crop yield prediction using hybrid machine learning approach: A case study of lentil (lens culinaris medik.). Agriculture **13**(3), 596 (2023)

[35] Akbar, A., Kuanar, A., Patnaik, J., Mishra, A., Nayak, S.: Application of artificial neural network modeling for optimization and prediction of essential oil yield in turmeric (curcuma longa l.). Computers and Electronics in Agriculture **148**, 160–178 (2018)

[36] Yildirim, T., Moriasi, D.N., Starks, P.J., Chakraborty, D.: Using artificial neural network (ann) for short-range prediction of cotton yield in data-scarce regions. Agronomy **12**(4), 828 (2022)

[37] Torkashvand, A.M., Ahmadi, A., Nikravesh, N.L.: Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ann) and multiple linear regressions (mlr). Journal of Integrative Agriculture **16**(7), 1634–1644 (2017)

[38] Khaki, S., Pham, H., Wang, L.: Yieldnet: A convolutional neural network for simultaneous corn and soybean yield prediction based on remote sensing data. bioRxiv (2020). 2020-12

[39] Alvarez, R.: Predicting average regional yield and production of wheat in the argentine pampas by an artificial neural network approach. European Journal of Agronomy **30**(2), 70–77 (2009)

[40] Li, A., Liang, S., Wang, A., Qin, J.: Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photogrammetric Engineering & Remote Sensing **73**(10), 1149–1157 (2007)

[41] Zhang, Y., Zhao, Z., Zheng, J.: Catboost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of northern china. Journal of Hydrology **588**, 125087 (2020)

[42] Leroux, L., Castets, M., Baron, C., Escorihuela, M.J., Bégué, A., Seen, D.L.: Maize yield estimation in west africa from crop process-induced combinations of multi-domain remote sensing indices. European Journal of Agronomy **108**, 11–26 (2019)

[43] Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A., Wajid, S.A.: Forecasting wheat yield from weather data and modis ndvi using random forests for punjab province, pakistan. International Journal of Remote Sensing **38**(17), 4831–4854 (2017)

[44] Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F.: Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and Forest Meteorology **218**, 74–84 (2016)

Supplementary Files

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

- SouthSudan20182019VF.xlsx
- SSUDANVF20182021Yieldfinal.xlsx