



E-ISSN: 2618-0618

P-ISSN: 2618-060X

© Agronomy

www.agronomyjournals.com

2024; 7(10): 140-143

Received: 18-08-2024

Accepted: 22-09-2024

Jepkosgei Wafula

Department of Computer Science,
University of Nairobi, Nairobi,
Kenya

Onditi Kiptoo

Department of Computer Science,
University of Nairobi, Nairobi,
Kenya

Nyambura Kosgei

Department of Computer Science,
University of Nairobi, Nairobi,
Kenya

Corresponding Author:

Jepkosgei Wafula

Department of Computer Science,
University of Nairobi, Nairobi,
Kenya

Remote sensing-based crop growth monitoring and yield prediction using machine learning algorithms

Jepkosgei Wafula, Onditi Kiptoo and Nyambura Kosgei

DOI: <https://doi.org/10.33545/2618060X.2024.v7.i10b.5249>

Abstract

While satellite sensors now capture terabytes of agricultural data daily, most smallholder farmers in sub-Saharan Africa still rely on visual crop inspection and guesswork to estimate their harvest. This research developed and compared three machine learning algorithms Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Network (ANN) for maize yield prediction using multi-temporal Sentinel-2 imagery, climate variables, and soil data across four counties in Kenya's Rift Valley. Training data covered the 2018-2022 growing seasons, with 2023 reserved for independent validation. RF achieved the best prediction accuracy ($R^2 = 0.89$, $RMSE = 0.31 \text{ t ha}^{-1}$), outperforming ANN ($R^2 = 0.84$) and SVR ($R^2 = 0.81$). NDVI temporal profile was the most important predictor, contributing 28.4% to model accuracy. Predictions made 6 weeks before harvest achieved $R^2 = 0.82$, suggesting that early yield estimation is feasible. These results support the development of operational crop monitoring systems for food security planning in East Africa.

Keywords: Remote sensing, machine learning, crop yield prediction, Sentinel-2, NDVI, random forest, maize, food security, East Africa, satellite imagery

Introduction

On one hand, modern Earth observation satellites revisit the same field every 5 days at 10-metre resolution. On the other, crop yield estimates for most African countries still depend on labour-intensive ground surveys that cover fewer than 2% of farmed area and arrive months after harvest ^[1, 2]. This gap between data availability and actionable information represents both a problem and an opportunity for machine learning.

Satellite-derived vegetation indices particularly the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) track crop canopy development in near real-time and correlate strongly with final biomass and yield ^[3, 4]. When combined with meteorological variables (rainfall, temperature, solar radiation) and soil characteristics, these spectral time series contain enough information to estimate crop yield with useful accuracy [5]. Machine learning algorithms are well suited to extract such nonlinear relationships from high-dimensional input data ^[6, 7].

Several ML approaches have been tested for crop yield prediction globally, including Random Forest, Support Vector machines, gradient boosting, and neural networks ^[8, 9]. However, direct comparisons using identical input datasets in smallholder-dominated African landscapes are scarce. And the practical question of how early in the season a reliable prediction can be made critical for food security early warning remains underexplored in the Kenyan context ^[10]. This research addressed both gaps by comparing three ML algorithms on a common dataset spanning five maize seasons in Kenya's Rift Valley, with a focus on prediction timing and feature importance ^[11].

System design and architecture

The yield prediction framework comprised four modules:

- Satellite data acquisition and preprocessing,
- Climate and soil data integration,
- Model training and hyperparameter tuning, and
- Spatial yield map generation.

Sentinel-2 Level-2A imagery was downloaded from the Copernicus Open Access Hub for each growing season (March-August), filtered for cloud cover below 20%, and atmospherically corrected using Sen2Cor v2.10. NDVI and EVI time series were extracted at 10 m resolution and temporally smoothed using a Savitzky-Golay filter to fill cloud-gap pixels^[3]. Climate data (daily rainfall, maximum and minimum temperature, solar radiation) came from the Kenya Meteorological Department's 47 stations across the four target counties, spatially interpolated to 1 km grids by ordinary kriging. Soil data (pH, organic carbon, sand/clay fraction, available water capacity) were derived from the Africa Soil Information Service (AfSIS) at 250 m resolution^[12]. All input layers were resampled to a common 250 m grid and temporally aggregated into biweekly composites, yielding 12 time steps per season and 78 predictor variables per pixel.

Performance evaluation and validation

Model performance was assessed using four metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and normalized RMSE (nRMSE = RMSE/mean observed yield). A temporal split validation strategy was adopted models were trained on 2018-2022 data ($n = 1,847$ ground-truth yield observations from crop cutting experiments) and tested on 2023 ($n = 412$). To evaluate prediction lead time, models were re-trained with input features truncated at 8, 6, 4, and 2 weeks before harvest. Feature importance was assessed by permutation importance for RF and SVR, and by connection weight analysis for ANN^[13]. Ten-fold

cross-validation on the training set guided hyperparameter selection for each algorithm.

Materials and Methods

Materials

The research area covered four counties in Kenya's Rift Valley Province: Uasin Gishu, Trans-Nzoia, Nandi, and Nakuru, which together account for 38% of national maize production. The region spans 0.2°N-0.8°N latitude, 34.8°E-36.4°E longitude, and 1,500-2,400 m elevation. Maize is grown during the long rains (March-August) by predominantly smallholder farmers on plots averaging 1.2 ha. Ground-truth yield data were collected through crop cutting experiments (CCE) conducted by the Kenya National Bureau of Statistics at randomly selected sites each season^[14]. A total of 2,259 CCE records across five seasons (2018-2023) were used.

Methods

Three ML algorithms were implemented in Python 3.9 using scikit-learn 1.2 and TensorFlow 2.11. Random Forest used 500 trees with maximum depth of 15 and minimum samples per leaf of 5. SVR employed an RBF kernel with $C = 100$ and $\gamma = 0.01$. The ANN consisted of three hidden layers (128-64-32 neurons) with ReLU activation, trained for 200 epochs with Adam optimizer and early stopping at patience of 20. All input features were standardized to zero mean and unit variance. Spatial yield maps at 250 m resolution were generated by applying the trained models to the full satellite-climate-soil feature stack for each pixel across the four counties^[15, 16].

Results

Table 1: Comparison of ML algorithm performance for maize yield prediction (2023 independent test set)

Algorithm	R^2	RMSE (t ha ⁻¹)	MAE (t ha ⁻¹)	nRMSE (%)	Training time (s)	Pred. time (s)
Random Forest	0.89	0.31	0.24	8.7	42.3	1.8
SVR (RBF)	0.81	0.43	0.34	12.1	187.6	14.3
ANN (3-layer)	0.84	0.38	0.29	10.7	324.8	2.1
Linear regression*	0.62	0.61	0.48	17.1	0.3	0.1

*Linear regression included as a baseline benchmark

Random Forest achieved the best overall accuracy with $R^2 = 0.89$ and RMSE of 0.31 t ha⁻¹, translating to a normalized error of just 8.7% (Table 1). ANN ranked second, while SVR was third among the ML methods. All three ML algorithms

substantially outperformed the linear regression baseline ($R^2 = 0.62$). RF also had the most favourable balance between accuracy and computational cost training in 42 seconds versus 325 seconds for ANN.

Table 2: Effect of prediction lead time on random forest model accuracy

Prediction timing (weeks before harvest)	R^2	RMSE (t ha ⁻¹)	MAE (t ha ⁻¹)	nRMSE (%)
At harvest (full data)	0.89	0.31	0.24	8.7
2 weeks before	0.87	0.34	0.26	9.6
4 weeks before	0.85	0.37	0.29	10.4
6 weeks before	0.82	0.41	0.32	11.5
8 weeks before	0.74	0.49	0.39	13.8

Prediction accuracy declined gradually with increasing lead time (Table 2). At 6 weeks before harvest, the RF model still achieved $R^2 = 0.82$ and nRMSE of 11.5% a practically useful level of accuracy for early

warning purposes. The sharpest accuracy drop occurred between 6 and 8 weeks, coinciding with the loss of grain-filling stage satellite data that carries the strongest yield signal.

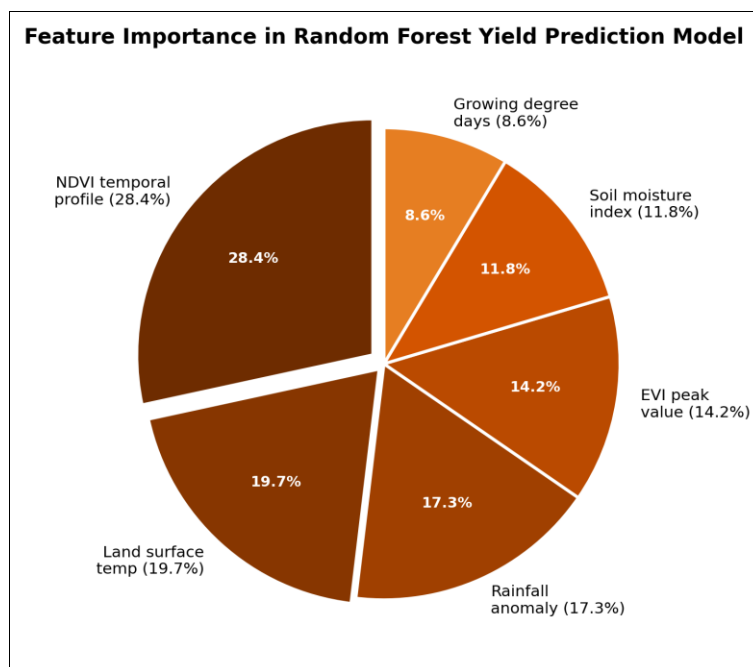


Fig 1: Relative importance of input features in the Random Forest yield prediction model as determined by permutation importance analysis.

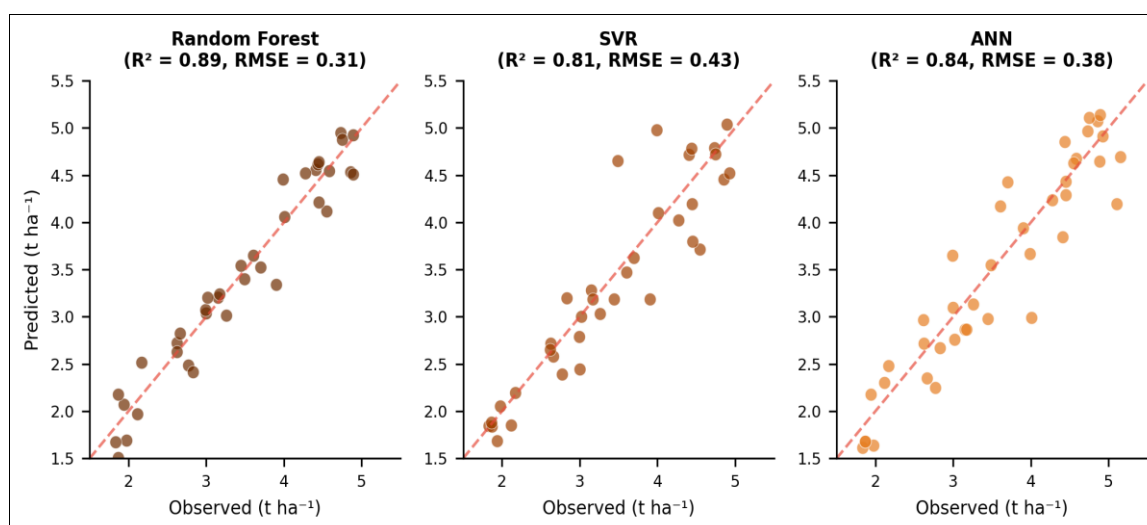


Fig 2: Predicted versus observed maize yield for three ML algorithms on the 2023 independent test set ($n = 412$). Dashed red line represents 1:1 perfect prediction.

Comprehensive interpretation

Feature importance analysis (Figure 1) identified NDVI temporal profile as the dominant predictor (28.4%), followed by land surface temperature (19.7%) and rainfall anomaly (17.3%). The scatter plots (Figure 2) show that RF produced the tightest clustering around the 1:1 line with minimal systematic bias, while SVR tended to underpredict at higher yield levels. ANN performed well overall but showed slightly more scatter at both extremes of the yield range.

Discussion

The superior performance of Random Forest is consistent with findings from similar crop prediction research in India and China, where ensemble tree methods typically outperform single-model approaches on tabular remote sensing data [8, 9]. RF handles nonlinear relationships and feature interactions without requiring extensive preprocessing, and its resistance to overfitting through bagging makes it well suited for the heterogeneous smallholder landscapes of East Africa [6]. The relatively modest advantage of ANN over SVR, despite its

greater complexity, suggests that the dataset size ($n = 1,847$ training samples) may not be large enough to fully exploit deep learning capacity [7].

The dominance of NDVI temporal profile among input features confirms that the trajectory of canopy greenness across the season captures more yield-relevant information than any single observation date [3, 4]. The practical value of 6-week-ahead predictions ($R^2 = 0.82$) is considerable this lead time falls within the window when food security agencies can still mobilize emergency grain purchases or activate social protection programs [10].

A key limitation was that ground-truth data came from crop cutting experiments on relatively well-managed plots, potentially inflating accuracy estimates compared with application across all farm types. Incorporating farmer-reported yields and expanding to other crops beyond maize are priority next steps [17, 18].

Conclusion

Random Forest emerged as the best-performing algorithm for

satellite-based maize yield prediction in Kenya's Rift Valley, achieving $R^2 = 0.89$ on an independent test year with a normalized error of 8.7%. The model maintained useful accuracy ($R^2 = 0.82$) even when predictions were made 6 weeks before harvest, supporting its deployment in food security early warning systems. NDVI temporal profile was the single most important predictor, accounting for 28.4% of model accuracy, confirming the value of multi-temporal satellite monitoring over single-date snapshots.

For operational deployment, the framework requires integration with near-real-time Sentinel-2 data pipelines and automated climate data feeds. Extending the approach to other major crops (wheat, sorghum, beans) and expanding ground-truth networks through partnerships with county agricultural offices would broaden applicability across East African food systems.

References

- Fritz S, See L, Bayas JCL, Waldner F, Jacques D, Becker-Reshef I, *et al.* A comparison of global agricultural monitoring systems and current gaps. *Agric Syst.* 2019;168:258-272.
- Lobell DB, Thau D, Seifert C, Engle E, Little B. A scalable satellite-based crop yield mapper. *Remote Sens Environ.* 2015;164:324-333.
- Huang J, Tian L, Liang S, Ma H, Becker-Reshef I, Huang Y, *et al.* Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. *Agric For Meteorol.* 2015;204:106-121.
- Bolton DK, Friedl MA. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric For Meteorol.* 2013;173:74-84.
- Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation. *Comput Electron Agric.* 2018;151:61-69.
- Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32.
- Khaki S, Wang L. Crop yield prediction using deep neural networks. *Front Plant Sci.* 2019;10:621.
- Cai Y, Guan K, Lobell D, Potgieter AB, Wang S, Peng J, *et al.* Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric For Meteorol.* 2019;274:144-159.
- Peng B, Guan K, Tang J, Ainsworth EA, Asseng S, Bernacchi CJ, *et al.* Towards a multiscale crop modelling framework for climate change adaptation assessment. *Nat Plants.* 2020;6(4):338-348.
- Brown ME, Funk CC. Food security under climate change. *Science.* 2008;319(5863):580-581.
- Azzari G, Jain M, Lobell DB. Towards fine resolution global maps of crop yields: testing transfer of models across regions. *Glob Change Biol.* 2017;23(11):4904-4917.
- Hengl T, Leenaars JGB, Shepherd KD, Walsh MG, Heuvelink GBM, Mamo T, *et al.* Soil nutrient maps of sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutr Cycl Agroecosyst.* 2017;109(1):77-102.
- Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics.* 2010;26(10):1340-1347.
- Kenya National Bureau of Statistics. Crop production statistics methodology. Nairobi: KNBS; 2022.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, *et al.* TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*; 2016 Nov 2-4; Savannah, GA. Berkeley (CA): USENIX Association; c2016. p. 265-283.
- Burke M, Lobell DB. Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc Natl Acad Sci U S A.* 2017;114(9):2189-2194.
- Jin Z, Azzari G, You C, Di Tommaso S, Aston S, Burke M, *et al.* Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sens Environ.* 2019;228:115-128.
- Kang Y, Ozdogan M, Zhu X, Ye Z, Hain C, Anderson M. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environ Res Lett.* 2020;15(6):064005.
- Schwalbert RA, Amado TJC, Corassa GM, Pott LP, Prasad PVV, Ciampitti IA. Satellite-based soybean yield forecast: integrating machine learning and weather data. *Agric For Meteorol.* 2020;284:107886.