Crop Harvest Prediction Using Remote Sensing and Machine Learning: An Integrated Framework

Deepanshu Singh[1], Kartikey Tiwari[2]
[1, 2]School of Information and Communication Technology, Gautam Buddha University, Greater Noida, UP, India

Abstract

Accurate estimation of the optimal harvest date (OHD) is critical for maximising crop yields and grain quality. This paper presents a machine learning pipeline that uses real soil and climate agronomic data alongside simulated Sentinel-2 Normalised Difference Vegetation Index (NDVI) time-series to classify crop types and compute an Extended Harvest Readiness Index (HRI) for four major crops: Maize, Wheat, Rice, and Cotton. NDVI time-series are modelled using the double-logistic phenology model (Zhang et al., 2003), with parameters derived from published Sentinel-2 field studies (Zhong et al., 2014; Xu et al., 2019). Five NDVI-derived features are extracted per field sample: peak NDVI, day of peak NDVI, NDVI at harvest date, season-integrated NDVI, and the rate of NDVI decline during senescence. Four classifiers are evaluated — Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM with RBF kernel), and Multilayer Perceptron (MLP) — on both tabular-only and NDVI-augmented feature sets. The best-performing model, Random Forest, achieves 100% test accuracy on tabular features and maintains this accuracy with NDVI augmentation. The Extended HRI combines model confidence, temperature suitability, and NDVI-based senescence indicators into a single actionable readiness score. Key findings confirm that NDVI characteristics — particularly ndvi_peak and ndvi_peak_doy — rank among the most important features, supporting the value of remote sensing augmentation for precision agriculture.

Keywords: crop classification, remote sensing, ndvi phenology, harvest readiness, machine learning, random forest, sentinel-2

1. Introduction

1.1 Problem Statement

Crop harvest timing is one of the most consequential decisions in modern agriculture. Xu et al. (2019) demonstrated that both premature and delayed harvest of maize leads to significant losses in yield and quality. Early harvest results in excessive kernel moisture, which promotes mildew and potentially carcinogenic aflatoxin formation. Delayed harvest exposes grain to field deterioration, frost damage, and quality degradation that occurs after physiological maturity.

Published: 08 May 2026
DOI: https://doi.org/10.70558/SPIJSH.2026.v3.i5.45726

Conventional harvest readiness indicators — including corn kernel moisture (CKM), milk-line position, and black-layer formation — depend on destructive field sampling techniques that do not scale effectively to large agricultural environments. Remote sensing offers a scalable, non-destructive alternative, enabling near real-time monitoring of canopy biochemical and structural changes across vast areas.

1.2 Research Objectives

This study pursues the following research objectives:

- Implement and benchmark four supervised classifiers for crop type identification on the real Crop Recommendation Dataset.
- Augment the tabular dataset with five NDVI-derived phenology features simulated using the double-logistic model.
- Measure the effect of remote sensing augmentation on classification accuracy.
- Develop an Extended HRI that integrates model confidence, temperature appropriateness, and NDVI senescence indicators.
- Deliver a fully reproducible Python pipeline deployable on commodity hardware.

2. Related Work

The methodological foundation of this study is drawn from Xu et al. (2019), who estimated canopy chlorophyll content (CCC) using HJ-1 CCD satellite imagery combined with the PROSAIL radiative transfer model, and established a non-linear relationship between CCC and CKM ($R^2 = 0.92$). Their approach determined OHD when CKM dropped below 30%, achieving a mean absolute prediction error of less than 1.3 days using only two satellite acquisitions.

Zhong et al. (2014) calibrated double-logistic phenological parameters for various crops using Landsat time-series, demonstrating crop-specific spectral signatures distinguishable at field scale. These findings were extended by Cai et al. (2018) using Sentinel-2 data, which enabled sub-field phenological mapping with higher spatial resolution. Meng et al. (2015) combined NDVI and the Normalised Difference Water Index (NDWI) to estimate soybean OHD, while Zhang et al. (2003) originally proposed the double-logistic NDVI model employed in this work for MODIS-based vegetation phenology monitoring.

3. Data and Methodology

3.1 Dataset

This study uses the Crop Recommendation Dataset (Kaggle/UCI), a real agronomic benchmark containing 2,200 samples spanning 22 crop types. The dataset is filtered to four crops — Maize, Wheat, Rice, and Cotton — resulting in 800 balanced samples (200 per crop). Each sample contains seven soil and climate features as described in Table 1.

Feature	Unit	Agronomic Meaning

N	mg/kg	Nitrogen content in soil
P	mg/kg	Phosphorus content in soil
K	mg/kg	Potassium content in soil
Temperature	°C	Average ambient temperature
Humidity	%	Relative humidity
pH	0–14	Soil pH
Rainfall	mm	Annual rainfall

Table 1. Tabular agronomic features used in this study.

3.2 NDVI Feature Engineering

NDVI time-series are simulated using the double-logistic phenology model proposed by Zhang et al. (2003):

$$NDVI(t) = NDVImin + (NDVImax - NDVImin) \times [1/(1 + exp(-a_1(t - b_1))) - 1/(1 + exp(-a_2(t - b_2)))]$$

Model parameters are estimated from Zhong et al. (2014) and Xu et al. (2019), with per-field Gaussian noise ($\sigma = 0.02$) added to simulate Sentinel-2 radiometric uncertainty. Five features are extracted from each simulated curve, as described in Table 2.

NDVI Feature	Remote Sensing Meaning	Harvest Relevance
ndvi_peak	Max canopy greenness	Higher peak indicates greater biomass potential
ndvi_peak_doy	DOY of peak NDVI (heading proxy)	Earlier peak suggests earlier possible harvest
ndvi_at_harvest	NDVI at harvest date (senescence)	Lower value indicates greater senescence and readiness
ndvi_integral	Cumulative season NDVI (biomass proxy)	Higher integral reflects a longer productive season
ndvi_sen_rate	NDVI decline rate, last 30 days	Faster rate indicates more rapid senescence

Table 2. NDVI-derived features and their agronomic interpretation.

3.3 Machine Learning Models

Four classifiers are trained on two feature configurations — tabular only and tabular augmented with NDVI features — using 5-fold stratified cross-validation. Model configurations are summarised in Table 3.

Model	Hyperparameters	Scaling
Random Forest	n_estimators = 200	None
Gradient Boosting	n_estimators = 150, lr = 0.1, depth = 4	None
SVM (RBF)	C = 10, gamma = scale	StandardScaler
Neural Network	layers = (128, 64, 32), ReLU, early stop	StandardScaler

Table 3. Model configurations. All models evaluated via 5-fold stratified cross-validation.

3.4 Extended Harvest Readiness Index (HRI)

The Extended HRI combines three complementary signals inspired by the CKM-threshold OHD model of Xu et al. (2019):

$$HRI = 0.50 \times P(class) + 0.25 \times Temperature\ Score + 0.25 \times NDVI\ Score$$

Where P(class) represents the highest predicted class probability (model confidence). The Temperature Score applies a Gaussian penalty for deviations from the crop-specific optimal temperature range. The NDVI Score is computed as: 0.6 × (1 − ndvi_at_harvest / 0.5) + 0.4 × (ndvi_sen_rate / 0.02), encoding the degree of canopy senescence. Decision thresholds are: HRI ≥ 0.80 indicates Ready to Harvest; HRI ≥ 0.60 indicates Near Ready; HRI < 0.60 indicates Not Ready.

4. Results

4.1 Exploratory Data Analysis — Feature Distributions

Feature distribution analysis confirms that the four crops occupy distinctly separated regions in feature space, validating the agronomic meaningfulness of the dataset and providing theoretical justification for high classifier accuracy. Low soil nitrogen and low temperature are the primary discriminators for wheat. Cotton exhibits the highest potassium values among all four crops. Rice requires substantially higher rainfall than the other three crops. These clear separations strongly support the potential for accurate classification.

4.2 NDVI Phenology Curves

Simulated Sentinel-2 NDVI curves are plotted using the double-logistic model for all four crops. Maize reaches peak NDVI at day of year (DOY) 201 with harvest at DOY 270. Wheat, which grows in the spring season, achieves its earliest peak at DOY 131 and harvest at DOY 195. Rice and Cotton both have late-season harvest dates at DOY 305 and DOY 315,

respectively. Shaded bands representing the P10–P90 field variability ensemble directly reflect inter-field uncertainty as described by Xu et al. (2019).

4.3 Feature Correlations with Harvest DOY

Pearson correlation analysis of all twelve features against Harvest DOY reveals that the strongest positive predictors are humidity ($r = 0.777$) and ndvi_peak_day ($r = 0.669$), while the strongest negative predictors are ndvi_peak ($r = -0.645$) and ndvi_at_harvest ($r = -0.424$). This biological relationship holds that crops maturing earlier exhibit earlier peak senescence dates, directly supporting the underlying agronomic hypothesis. The NDVI features thus provide biologically meaningful directional information about crop harvest timing.

4.4 Model Performance

Table 4 presents the 5-fold cross-validation and test accuracy of all four classifiers across both feature configurations. Random Forest and Gradient Boosting achieve near-perfect accuracy on tabular features alone (1.000 and 0.992 CV accuracy, respectively), consistent with the strong class separability observed in the exploratory analysis. SVM accuracy improves by +0.017 with NDVI augmentation. The Neural Network experiences a slight accuracy decline of −0.017 when NDVI features are added, attributable to increased feature dimensionality relative to the available training data.

Model	CV Acc (Tabular)	CV Acc (+ NDVI)	Test Acc (+ NDVI)	Δ Accuracy
Random Forest	1.000	1.000	1.000	0.000
Gradient Boosting	0.992	0.992	0.992	0.000
SVM (RBF)	0.983	1.000	1.000	+0.017
Neural Network	0.992	0.975	~0.975	−0.017

Table 4. Full model performance summary across both feature sets.

4.5 Feature Importance

Analysis of Random Forest feature importances (Mean Decrease in Impurity) reveals that among tabular features, humidity (0.2251), nitrogen — N (0.2013), rainfall (0.1928), and potassium — K (0.1600) are the dominant predictors. Notably, ndvi_peak (0.0892) and ndvi_peak_doy (0.0706) rank 5th and 6th overall, demonstrating that remote sensing phenology biomarkers provide discriminative value beyond soil chemistry alone. Senescence-related features (ndvi_at_harvest, ndvi_sen_rate) contribute less to crop classification but serve as essential inputs to the HRI calculation.

4.6 Confusion Matrix

The confusion matrix for the best model (Random Forest with full features) shows perfect

classification for Cotton, Maize, and Rice (20/20 each on the test set), demonstrating that agronomic and phenological signals are sufficient to differentiate between these crops without error. Wheat samples do not appear in the test set due to a sorting-based data merge artefact that clustered all wheat samples into the training fold; this limitation is addressed in the Discussion section.

5. Discussion

5.1 Alignment with Xu et al. (2019)

The foundational contribution of Xu et al. (2019) was the integration of canopy biochemistry (CCC) with agronomic maturity indicators (CKM) and satellite imagery to forecast OHD at scale. This reasoning is directly reflected in the HRI architecture of the present study: ndvi_at_harvest represents the current degree of senescence (analogous to measured CKM), ndvi_sen_rate encodes the velocity of canopy decline (analogous to the 0.54% per day CKM drop rate reported by Xu et al.), and the 30% CKM harvest threshold corresponds to the HRI ≥ 0.80 Ready threshold adopted here.

The observed negative correlation between ndvi_peak and Harvest DOY ($r = -0.645$) is particularly noteworthy: crops with earlier peak greenness dates also tend to mature earlier (e.g., Wheat at DOY 131 with harvest at DOY 195), while late-maturing crops such as Rice peak at DOY 233 and harvest at DOY 305. The double-logistic simulation successfully captures this cross-crop phenological differentiation, confirming its utility as a proxy for real Sentinel-2 observations.

5.2 Limitations

Several limitations constrain the current study. First, NDVI time-series are simulated rather than observed; integration with Google Earth Engine would provide spatially explicit, field-level phenological data inclusive of real-world atmospheric and cloud effects. Second, all wheat samples were inadvertently placed in the training set due to a sorting artefact during the label merge step, which is remedied by adopting a random stratified split. Third, the Neural Network's accuracy decline with NDVI features suggests insufficient data for the higher-dimensional input space, requiring either additional samples or enhanced regularisation. Finally, the CCC–CKM relationship established by Xu et al. (2019) is site-specific, implying that HRI thresholds require local calibration before operational deployment.

5.3 Future Work

Future research should incorporate real Sentinel-2 time-series via the Google Earth Engine API to replace simulated NDVI with observed field data. Additional spectral indices suggested by Xu et al. (2019) — including the Red-Edge Chlorophyll Index (CI_7) and NDWI — should be integrated as supplementary remote sensing features. Growing-degree-day accumulation coupled with established crop growth models such as DSSAT or APSIM could further strengthen the HRI formulation. A fully random stratified train/test split should be implemented to ensure all four crops are represented proportionally in both data partitions.

6. Conclusion

This study presents and validates an end-to-end machine learning framework for crop classification and harvest readiness prediction across four major crops. The framework bridges tabular agronomic data with remote sensing phenological signals by simulating Sentinel-2 NDVI time-series using the double-logistic model and extracting five physiologically grounded features. The Random Forest classifier achieves 100% test accuracy on the full 12-feature set. Key NDVI features — ndvi_peak (importance: 0.089) and ndvi_peak_doy (importance: 0.071) — rank among the top six most important predictors, validating the utility of remote sensing augmentation for crop-type discrimination.

Feature correlation analysis confirms the biological validity of the approach: ndvi_at_harvest and ndvi_sen_rate are negatively correlated with Harvest DOY ($r = -0.42$ and $r = -0.49$, respectively), directly mirroring the CKM decline mechanism described by Xu et al. (2019). The Extended HRI provides a multidimensional, actionable harvest decision score that can function as a scalable complement to destructive field sampling, advancing the vision of precision agriculture at regional scale.

References

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment, 201, 231–245.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Daughtry, C. S. T., Walthall, C. L., Kim, M. S., de Colstoun, E. B., & McMurtrey, J. E. (2000). Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sensing of Environment, 74(2), 229–239.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25–36.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sensing of Environment, 58(3), 289–298.

Immitzer, M., Vuolo, F., & Atzberger, C. (2016). First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sensing, 8(3), 166.

Jacquemoud, S., & Baret, F. (1990). PROSPECT: A model of leaf optical properties spectra. Remote Sensing of Environment, 34(2), 75–91.

Jonsson, P., & Eklundh, L. (2004). TIMESAT — a program for analyzing time-series of satellite sensor data. Computers & Geosciences, 30(8), 833–845.

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778–782.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Meng, J., Xu, J., & You, X. (2015). Optimizing soybean harvest date using HJ-1 satellite imagery. Precision Agriculture, 16(5), 567–582.

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems Engineering, 114(4), 358–371.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. NASA Special Publication, 351, 309–317.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8(2), 127–150.

Verhoef, W. (1984). Light scattering by leaf layers with application to canopy reflectance modelling: The SAIL model. Remote Sensing of Environment, 16(2), 125–141.

Xu, J., Meng, J., & Quackenbush, L. J. (2019). Use of remote sensing to predict the optimal harvest date of corn. Field Crops Research, 236, 1–13.

Zhang, X., Friedl, M. A., Schaaf, C. B., Strahler, A. H., et al. (2003). Monitoring vegetation phenology using MODIS. Remote Sensing of Environment, 84(3), 471–475.

Zhong, L., Gong, P., & Biging, G. S. (2014). Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. Remote Sensing of Environment, 140, 1–13.