ISSN 2354-9114 (online), ISSN 0024-9521 (print)
Indonesian Journal of Geography Vol 58, No.2 (2026): 195-209
DOI: 10.22146/ijg.111239 website: https://jurnal.ugm.ac.id/ijg

INDONESIAN JOURNAL OF GEOGRAPHY

RESEARCH ARTICLE

A Data-Efficient Rice Yield Classification Framework Combining Hybrid PCA-RFE and Regularized Random Forest

Yagus Wijayanto[1], Rika Nurjannah[1], Maya Wenlow Saragih[1], Ika Purnamasari[1], Tri Wahyu Saputra[1], Suci Ristiyana[1], Ummi Sholikhah[1], Rachmat Abdul Gani[2]

[1]Faculty of Agriculture, Universitas Jember, Indonesia
[2]National Research and Innovation Agency, Research Organization for Agriculture and Food, Research Center for Food Crops, Kawasan Sains dan Teknologi (KST) Dr. (H.C) Ir. H. Soekarno

Received: 2025-09-15
Revised: 2026-06-26
Accepted: 2026-07-28
Published: 2026-08-05

Key words: PCA; RFE; Random Forest; Rice Yield Classification; Spatial Uncertainty

Correspondent email : yaguswijayanto001.faperta@unej.ac.id

Abstract. Accurate rice yield prediction by remote sensing data and machine learning is still a big challenge in limited resources of field survey where the sample size is often very small. This study tackles the core challenge of small sample size (n=39) in rice yield classification by proposing a hybrid feature engineering framework that combines Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) in a regularized Random Forest (RF) classifier. The methodology is based on 30 spectral features from multi-temporal Sentinel-2A imagery (15 spectral bands i.e. Bands 2, 3, 4, 5 and 8 for three acquisition dates and 15 vegetation indices). PCA revealed three principal components that explained approximately 90% of the total variance, while RFE selected the five most discriminative spectral features: Band 8 (Jan 24, 2025; Feb 18, 2025), GNDVI (Jan 24, 2025), NDVI (Jan 24, 2025), and SAVI (Feb 18, 2025). These were fused in a compact eight-dimensional hybrid feature space. Model evaluation was conducted using repeated stratified cross-validation (5-fold x 10 repeats) and an independent test set (holdout 30%). The average cross-validation accuracy of the hybrid model was 83.36%, ROC-AUC 0.9304, independent test accuracy 91.67% and Cohen's Kappa 0.8333. There was no over-fitting of the training-validation gap to 0.10 in the learning curve. The optimal classification threshold was identified as 0.5373 in the Youden Index optimization. The RFE-selected spectral bands contributed the largest share (76%) in the feature importance analysis, and the PCA components provided a complementary 24%, confirming the synergistic value of unsupervised extraction and supervised selection. However, spatial uncertainty mapping revealed limitations to the overall extrapolation, with near-maximum values (0.999) in unsampled areas, emphasizing the need for targeted field verification in high-uncertainty zones. The study provides a replicable methodological blueprint for crop yield classification under very limited data conditions and provides practical guidance for agricultural monitoring in developing countries with logistical and financial constraints.

1. Introduction

Rice (Oryza sativa L.) is an important staple food. It is the major source of calories for more than half of the global population. Rice is an important crop for economic stability and food security (Mohidem et al., 2022). The growth of the world population and the rising demand for food have put tremendous stress on the production systems (Lam, 2024 ; Nwamekwe, 2025). Thus, the accuracy classification of rice yield categories is a vital in agricultural planning and decision-making Ashim et al., (2024) This process allows for optimal use of resources, improves crop management techniques and supports the development of effective policies to guarantee sustainable productivity, including food security (Jeong et al., 2022; Wang et al., 2024; Wei et al., 2023).

The development of remote sensing technology has greatly improved the efficiency of agricultural monitoring and has revolutionized the traditional ground-based methods by providing large-scale crop assessment (Guo et al., 2021;Khanal et al., 2020). The Sentinel-2A satellite is a state-of-the-art instrument that delivers high spatial resolution (10 m) multispectral data with an increased revisit frequency (every 5 days) and an open-data policy (Xie et al., 2025). The obtained data are important for deriving vegetation indices such as NDVI, SAVI and GNDVI, which are proxies for relevant crop health measures such as biomass and chlorophyll content, which are important for yield prediction (Kazemi & Parmehr, 2023; Imanni et al., 2022). Meanwhile, the application of remote sensing data in combination with machine learning techniques is now on the rise for classification and crop yield prediction purposes (Haque et al., 2025). In this regard, the combination of remote sensing, GIS and precision agriculture technologies effectively improves the crop monitoring and facilitates sustainable agriculture through the optimal use of irrigation, fertilizers and pesticides, thus reducing resource consumption and environmental degradation (Wang et al., 2024).

The sample size is often limited, which is a problem that hinders the optimization of machine learning applications.

In the case of agricultural geography, this problem is further aggravated by the fact that the collection of the ground truth data for model calibration and validation requires high operational costs, trained human resources and access to geographically difficult locations (Araújo et al., 2023). In this study, there were 39 samples, a small number for building classification models for tens or hundreds of features, which is a common situation in research. This problem is called the high dimensionality low sample size (HDLSS) problem. The main problem is the high tendency of the model to overfitting where the accuracy of training is high but generalization to new data is low (Henry & Stinchcombe, 2025). It is worth mentioning that similar challenges are faced in a number of other fields and researchers are still struggling with a problem of analyzing such data (Issa et al., 2024; Manzoor et al., 2024). Conventional classification approaches often fail when the number of features is much larger than the number of observations. For even small samples, K-fold cross-validation produces biased estimates of performance, which also holds for the increased sample size of 1000 (Vabalas et al., 2019).

A number of methods have been proposed to overcome the small sample problem in classification. An alternative to deep learning techniques is feature engineering, which comprises feature extraction and selection. Feature engineering does not require additional data to improve the classification performance, unlike deep learning that needs large datasets and is often subject to mode collapse in GAN-based data augmentation. The problem of high-dimensional data can be solved with the help of PCA (Principal Component Analysis) which is an effective method of reducing dimensionality (Nayana et al., 2022; Pasternak & Pawluszek-filipiak, 2022). PCA helps to transform a set of correlated variables into smaller uncorrelated components without losing the maximum variance of the bands (M. Chen et al., 2024). Moreover, PCA application helps to remove redundant information and preserve significant spectral features, thereby makes the data suitable for machine learning models. Simultaneously, Recursive Feature Elimination (RFE) has been introduced as a supervised feature selection approach that successively eliminates the least significant features based on the feature importance ranking of a model (Priyatno et al., 2024; Issa et al., 2024). PCA and RFE are common feature engineering methods used in machine learning pipelines, where PCA is an unsupervised technique to capture global variance, and RFE is a supervised technique to select locally discriminative features (Manzoor et al., 2024; Shin & Oh, 2024)

The selection of Random Forest (RF) as the central part of this hybrid PCA-RFE framework is not accidental, but rather due to its unique algorithmic structure which distinguishes it from Support Vector Machines (SVM), Neural Networks or linear models, particularly for the case of very small sample sizes. RF is a popular technique that has been widely used for its high accuracy, robustness to overfitting and ability to handle complex high dimension data without making any strong assumptions on the data distribution (Sah et al., 2024; Sun et al., 2025). RFE needs a supervised estimator for the feature importance ranking (Li et al., 2023). The RF provides very stable non-parametric measures of importance based on intrinsic Gini impurity reduction and out-of-bag (OOB) permutation tests, which can naturally capture complex non-linear interactions between spectral bands and rice yield (Moraiti et al., 2024). Also, linear models such as Logistic Regression and Linear SVM are unable to deal with non-linearity in the data. While a non-linear SVM kernel such as RBF can learn to represent a very non-linear boundary in the feature space, it is not easy to optimize the SVM hyperparameters C and gamma in high-dimension and low-sample-size settings. SVM also can easily capture noise in the data (Shang et al., 2025). PCA regularization can help reduce multicollinearity but some studies indicate that it cannot prevent feature masking which refer to the situation where a feature competing with other features at split selection (Gunduz & Fokoue, 2015) (Gunduz & Fokoue, 2015).

Ensemble model is less prone to overfit than its components. The algorithm finds locally discriminative bands by regularizing the feature space. PCA regularization reduces the global spectral variance by de-correlating the highly multicollinear vegetation indices and making the feature space orthogonal. In contrast, RFE performs local regularization of the feature space by eliminating redundant features. The ensemble model regularizes the effective dimensionality of the feature space as the number of samples decreases. The approach also differs from algorithms such as Neural Networks which rely on a diverse parameter space and are known to perform poorly when the number of samples is limited. The proposed approach is superior for the processing of HDLSS data as demonstrated by (Gunduz & Fokoue, 2015). The benefits of combining multiple feature engineering strategies have been well documented. For instance, Shin & Oh, (2024) showed that a hybrid feature selection method designed for HDLSS settings could significantly reduce the mean number of selected features from 37.8 to 5.5 and improve the prediction performance from 0.855 to 0.927. Their results confirm the fact that hybrid approaches often provide better results than any single technique, as also supported by the ensemble feature selection literature (Kamalov et al., 2023; Chen et al., 2023).

Previous studies on RF models have demonstrated their effectiveness in predicting crop yield, but there is still a large gap in knowledge on the best input features for models in highly heterogeneous agricultural environments such as smallholder farms (Elders et al., 2022). Vegetation indices have been successfully explored, however, few analyses have directly compared their performance with multispectral bands within PCA-based RF frameworks (Mohammadpour et al., 2022; Pham et al., 2022). However, the combination of PCA and RFE as a feature fusion structure to classify rice yield under small sample conditions has not been widely studied in the agricultural and geography literature.

In this paper, a hybrid structure is proposed, where PCA is used to extract uncorrelated features from the vegetation indices, and RFE is used to derive the most predictive subset of spectral characteristics. Then the outputs of the PCA transformation and RFE selection are combined and fed into the regularized RF classifier for classifying the target field into rice yield classes, minimizing the risk of overfitting on the ultra-small available dataset. The objective of this study is to fill this knowledge gap by performing a comparative analysis in order to identify models of rice yield classes based on two different sets of input features derived from Sentinel-2A bands and vegetation indices in a hybrid PCA-RFE-RF framework. Essentially, it is about determining which data source is a better representation of crop health and yields potential for predicting rice productivity in a smallholder farming setting, as well as validating the model from diverse perspectives.

Three research questions are addressed: (a) quantification of the classification accuracy on small samples (n=39), (b) spatial depiction of the uncertainty to provide a profile of the

spatial uncertainty for decision-support and (c) detection of the most informative spectral determinants of rice yield in the studied region. The novelty of the paper includes three points, 1) PCA and RFE integration as a feature fusion structure for rice yield classification with small samples condition (n=39), 2) PCA-RFE combination in rice yield classification in the setting of agricultural and geography literature, and 3) integration of PCA and RFE have not been explored widely in the above literatures. The study also provides additional accuracies in reporting through learning curve analysis for audit of overfitting, Youden Index analysis for threshold optimization, and spatial mapping for prediction of probabilities and uncertainty projection, to provide a more complete picture of the model and its reliability projection. This study combines feature extraction and selection, regularized classification and spatial uncertainty visualization tools, and offers a full pipeline that meets the demands of small-sample remote sensing projects and gives geography researchers practical guidance on how to set up machine learning experiments under non-ideal data conditions. This method works well even in data-scarce regions and can be applied elsewhere, serving as a replicable methodological blueprint for crop classification under very limited data conditions (Khan et al., 2024).

2. Methods

Study area and dataset description

This study employed secondary and primary data to develop and validate the classification models. The secondary data were 3-dates satellite images of Sentinel-2A Level 2A of the Copernicus Open Access Hub during the rice planting seasons of December 10, 2024; January 24 , 2025; and February 18, 2025. The selected bands were Band 2 (Blue), Band 3 (Green), Band 4 (Red), Band 5 (Red Edge) and Band 8 (Near-Infrared). The spatial resolution was 10m for all the bands except Band 5, which was 20m and further resampling into 10m resolution.

The main data used were ground truth data of rice productivity obtained from 39 sample points using the Quadrant method. Further, a 2.5 m x 2.5 m bamboo plot was used to harvest and weigh the paddy grain (GKP). This weight was then converted to Dry Milled Grain (GKG) using a conversion factor of 83.17% and categorized into three different productivity classes: Low and High (binary class). The classes were used as categorical target variables for the training and validation of the RF classification models.

The study area was collected with a total of 39 field sample points that were geo-referenced using a handheld GPS device for accurate spatial alignment with the corresponding satellite imagery. The rice yield was calculated at each sampling point and subsequently divided into two binary classes, low-yield (Class 0) and high-yield (Class 1), based on the median yield value of all the samples, thus providing an objective and data-driven approach. The data distribution obtained was fairly balanced with 18 samples in Class 0 and 21 samples in Class 1 and does not pose a high risk of severe class imbalance during model training.

Spectral data were extracted for each sample point from multi-temporal Sentinel-2A satellite images, resulting in 30 spectral features labelled from `SAMPLE_1` to `SAMPLE_30`. These features are systematically grouped into two categories: the first 15 features (from `SAMPLE_1` to `SAMPLE_15`) represent individual spectral bands (i.e., Bands 2, 3, 4, 5, and 8 for three acquisition dates), and the last 15 features (from `SAMPLE_16` to `SAMPLE_30`) represent vegetation indices calculated using standard biophysical formulas (e.g., NDVI, EVI, SAVI) to capture various aspects of crop vigour and canopy structure. The following are the features (as for instance : SAMPLE_16) and the corresponding date: (a) December 10, 2024; January 24, 2025; and February 18, 2025. Therefore, there are 15 features (then it is called SAMPLE) for these three dates from Sentinel 2A. They were Band 2, Band 3, Band 4, Band 5, Band 8. And there were also 15 vegetation indices for these three dates of Sentinel 2A: GNDVI (Green Normalized Difference Vegetation Index), NDRE (Normalized Difference Red Edge index), NDVI (Normalized Difference Vegetation Index), SAVI (Soil Adjusted Vegetation Index), and VARI (Visible Atmospherically Resistant Index) where all formulas for calculating these indices can be found in (Yan et al., 2025; Radoˇ et al., 2023)

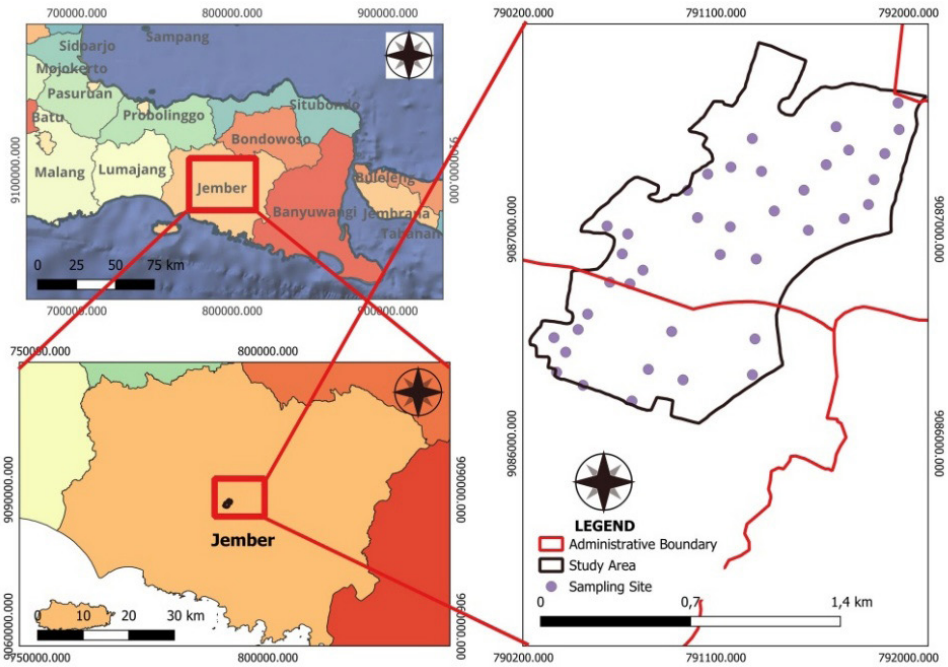

Figure 1. The study area and the locations of yield samples

Research framework and workflow

In this work, we adopted a systematic procedure consisting of unsupervised feature extraction using PCA, supervised feature selection using RFE and a regularized ensemble classifier (Random Forest). Firstly, the standardization and preprocessing the data was conducted. Next, we did feature engineering using PCA and RFE. We then proceeded to model training, with particular attention to tuning. Then evaluated performance evaluation was carried out using cross-validation and a separate test set. We also checked for any signs of overfitting. We finally visualized predictions, their probabilities and uncertainty. To justify the choice of 3 principal components, a scree plot was drawn to visualize the ratio of explained variance for each component. We did all processes in Python using Scikit-learn, and just to make sure everything is repeatable we set the random seed to 42.

Data pre-processing

To get rid of scale differences between spectral bands, we standardized the data before extracting features. We used StandardScaler separately on two data sets: first, all 30 features going into RFE, and second, the 15 vegetation indices set aside for PCA. Standardization set every feature to have a mean of zero and a variance of one. This keeps methods like PCA from getting skewed by features with bigger numerical ranges and makes sure RFE ranks features by importance instead of scale. This standardization preserves the relative relationships among bands while ensuring that no single feature dominates the PCA or RFE due to scale differences.

Hybrid feature engineering using RFE, PCA, and feature fusion

At the heart of this methodology is the synergistic combination of two different dimensionality reduction strategies:

Principal Component Analysis (PCA) was applied to the fifteen standardized vegetation indices to extract uncorrelated latent components that capture the global structure of the spectral variance. Based on the **scree plot**, the first two or three principal components were retained as these three components accounted for mostly the of the total variance in the vegetation indices. These components include information on vegetation density, chlorophyll content and canopy structure, but are not limited to one specific index. The following (Figure 2) is the scree plot of PCA from vegetation indices.

The blue bars and green bars show the percentage of variance explained by each principal component and the red dashed line shows the cumulative explained variance. The scree plot clearly exhibits a 'elbow' at the boundary between the third and fourth components (where the marginal gain in the explained variance drops sharply from 0.20 to 0.10). Therefore, we retained the first three principal components in the hybrid feature space which together account for about 85% of the total variance. The first three principal components capture the main spectral signals and are able to effectively exclude the residual noise represented by the later components. From the PCA loadings matrix (Table 1), it was observed that PC1 was highly influenced by SAMPLE_16 (GNDVI- December 10, 2024) and Sample_17 with loading values greater than 0.40

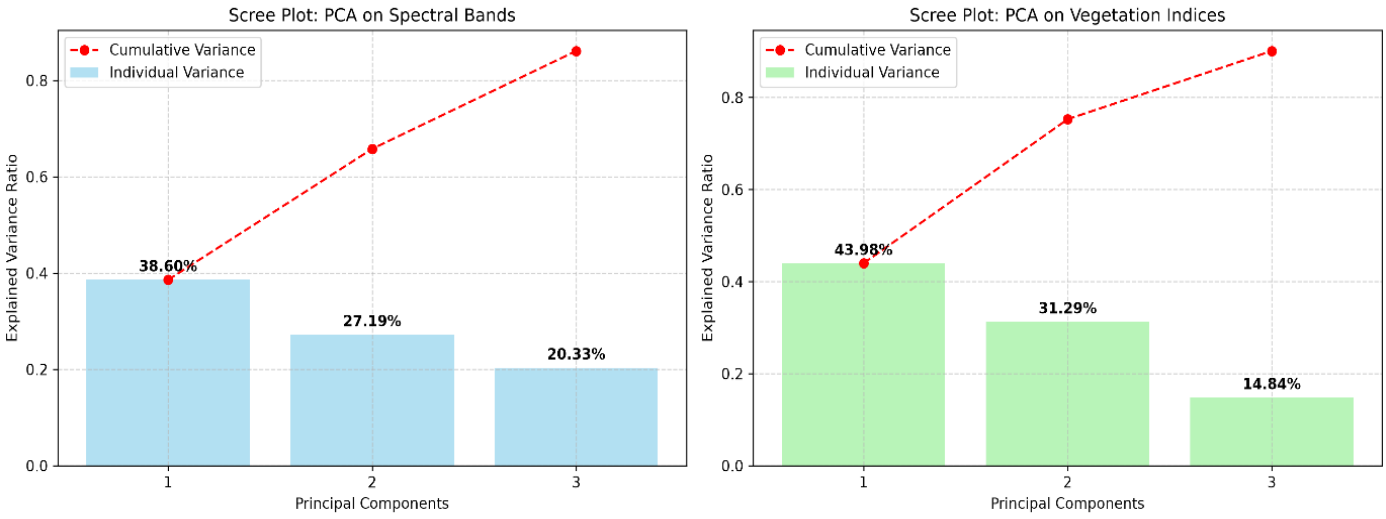

Figure 2. Scree plot of PCA for Spectral Bands and Vegetation indices

Table 1. PCA loadings for the first three Principal Components (Vegetation Indices)

Feature	PC1	PC2	PC3
SAMPLE_16 (GNDVI- December 10, 2024)	**0.42**	0.15	0.08
SAMPLE_17 (GDNVI-Januari 24, 2025)	**0.41**	0.12	0.10
SAMPLE_23 (NDVI-January 24, 2025)	0.18	**0.38**	0.14
SAMPLE_24 (NDVI-February 18, 2025)	0.16	**0.35**	0.12
SAMPLE_26 (SAVI-18 February, 2025)	0.20	0.18	**0.44**
SAMPLE_28 (VARI-December 10, 2024)	0.14	0.20	**0.38**

suggesting that these vegetation indices capture the primary spectral variance related to overall vegetation vigour and greenness. PC2 was mainly related to SAMPLE_23 (NDVI-January 24, 2025) and SAMPLE_24 (NDVI-February 18, 2025), which were related to additional variance associated with soil-adjusted vegetation signals and canopy structure. In contrast, PC3 explained variance related to SAMPLE_26 (SAVI-18 February, 2025) and SAMPLE_28 (VARI-December 10, 2024), indicating that this component captures subtle differences in chlorophyll content and canopy moisture, in addition to the coarser structural information captured by the first two components.

Recursive Feature Elimination (RFE) was performed on the full set of the 30 standardized spectral features. RFE uses a supervised estimator (here a Random Forest classifier with 50 trees) to rank the importance of each feature based on the decrease in Gini impurity. The features are recursively eliminated. The algorithm iteratively improves the subset to find the optimal set. This process selected the 5 most discriminative spectral bands (n_features_to_select=5): SAMPLE_14 (Band 8-January 24, 2025), SAMPLE_15 (Band 8-February 18, 2025), SAMPLE_17 (GNDVI-January 24, 2025), SAMPLE_23 (NDVI-January 24, 2025), SAMPLE_26 (SAVI-January 24, 2025). The selected features represent the most informative spectral variables in a local sense for discriminating low and high rice productivity classes. The feature fusion was done by horizontal concatenation (np.hstack) of the 3 principal components obtained from PCA and the 5 features selected by RFE. This process generates a hybrid feature matrix with 8 dimensions, which combines global structural information (PCA) and discriminative local spectral information (RFE) to provide a compact, information-rich and multi-perspective feature description for small-sample scenarios.

Random Forest Model Configuration and Regularization

In this study, we used a regularized Random Forest classifier as the base estimator to reduce the risk of overfitting on a limited dataset. A regularization strategy was used to constrain the hyperparameters, encouraging generalization over memorization: n_estimators=100 to ensure that the ensemble is sufficiently large for convergence; max_depth=3 to strictly restrict the complexity of the tree and prevent it from memorizing noise; min_samples_split=4 and min_samples_leaf=2, mainly to enforce minimum sample requirements for splits and leaf nodes while ensuring that splits are supported by enough data; bootstrap=True and oob_score=True by using bootstrap sampling to create diverse training subsets and using out-of-bag samples for internal validation. Meanwhile, we set class_weight='balanced' to adjust class weights automatically in inverse proportion to class frequencies to reduce any possible imbalance in the binary response variable, and we used n_jobs=-1 to use all available CPU cores to improve the computational efficiency. This setup allows the model to focus on generalization rather than memorization, which is an important characteristic when dealing with n=39 observations.

Training, testing and model evaluation

The hybrid feature matrix is then split into train and test by the stratified train-test split with a ratio of 70:30 (`test_size=0.3`). This stratification keeps the proportion of low and high outcome classes in both subsets, preventing biased performance estimates. Systematic evaluation was done based on two parallel approaches: (a) Independent test set evaluation on the 30% holdout data, reporting the accuracy, Cohen's Kappa and classification metrics (precision, recall, F1-score) and (b) Repeated stratified K-fold cross validation (5 folds, 10 repeats), with 50 different train-test splits to improve the robustness of performance estimates.

SVM and XGBoost were excluded from the comparison analysis due to their known limitations in HDLSS settings. Gunduz & Fokoue, (2015) demonstrate that the SVM, in particular with non-linear kernels, is prone to overfitting and is hyperparameter sensitive when the sample size is very small in comparison to the feature space. Besides, despite its good performance on larger datasets, XGBoost has been shown to require much more training samples to achieve stable convergence compared to Random Forest (C. Chen et al., 2023). The comparison with RF without PCA and CART as a simpler interpretable baseline was enough to isolate the contribution of the proposed hybrid approach, since the main aim of this work was to assess the effectiveness of feature compression strategies (PCA and RFE) in a regularized framework.

Overfitting analysis

The learning curves were used to check for model overfitting. The model was trained using incremental training set sizes (from 20% to 100% of the training data, in 5 steps). For each training size, the model performance (accuracy) is evaluated on the training subset and the cross-validation subset. The obtained learning curves would be used to evaluate the model. If a small and closing gap has occurred, then it is the indication of a well generalised model. If the gap is consistently wide (training accuracy much higher than validation accuracy), then it would be a case of severe overfitting. The diagnostic step gave an empirical basis for the reliability of the model in the next stage of spatial mapping. The learning curve was generated by incrementally increasing the size of the training set from 20% to 100% in five increments. At every step, the model was trained and evaluated using 5-fold cross-validation to obtain the training and validation accuracy. The obtained curves were used to evaluate the bias-variance trade-off and to identify potential overfitting.

ThThreshold optimization using Youden Index

In addition to the standard classification with the default probability threshold of 0.5, this study used the Youden Index to identify the optimal probability threshold for separating classes. The Youden Index (J = Sensitivity + Specificity - 1 ` or 'TPR – FPR') seeks the maximum value of the model to differentiate between classes by balancing the true positive rate (TPR) and false positive rate (FPR). The optimal decision cutoff is determined from the ROC curve generated on the test set, corresponding to the threshold that maximizes the Youden Index. Moreover, we perform a sensitivity analysis by defining two other tolerance thresholds, a positive adjustment of 8% and 15% of the optimal threshold. The best probability threshold was determined by the Youden Index, which finds the cutoff with the best discriminatory power of the model. Then the threshold was compared with the default threshold of 0.5 in order to examine the effect of threshold optimization on classification performance.

Spatial prediction

The new contribution of this methodology is the integration of spatial coordinates in the model output : (a) the

hybrid model predicts binary outcome classes for all sample locations, which are color-coded for a visual interpretation of the spread of productivity; (b) the underlying probabilistic output, highlighting areas with a high likelihood of high yield. This approach generates high-resolution 2D spatial visualizations using the 'X' and 'Y' coordinates for each sample. This produces a probability map; and (c) the extent of spatial uncertainty which can be calculated by the formula. The index ranged from 0 to 1. The mapping of this index allows to identify specific geographic zones with a 'lack of confidence' of the model or where the spectral information is not allowing a clear classification. These areas can then be targeted for further field verification or for the collection of alternative data.

Feature importance analysis

To understand the logic behind the decision making of the model, the Gini importance scores (feature importances) are taken from the trained Random Forest model. These scores are the contribution of each of the 8 hybrid features (3 PCs + 5 bands selected using RFE) to the classification task. Results are sorted and presented as bar charts. The result gave an indication whether the latent components (PCA) or the specifically selected spectral bands (RFE) were dominant in the prediction process. Feature importance was derived from the decrease in Gini impurity within the trained Random Forest model. The importance scores of the eight hybrid features (3 PCA components + 5 RFE-selected bands) were extracted using the feature_importances attribute and visualized through a bar chart to identify the dominant predictors.

3. Result and Discussion

Classification performance and model generalization

The classification performances of the Hybrid model are: (a) the mean CV accuracy is 83.36%, the mean CV ROC-AUC is 0.9304, and the test accuracy and Cohen's Kappa are 91.67% and 0.8333, respectively. The Test Accuracy of 91.67% is particularly interesting, as it is based on the 30% hold-out subset (12 samples) that was completely unseen. importantly this value is higher than the mean CV Accuracy (83.36%) which is actually a positive sign. Vabalas et al., (2019) warned that if feature selection (e.g. RFE) is done on the full dataset before splitting, the test set gets contaminated and leads to overly optimistic and biased results. In this study, strict methodological was kept: feature extraction (PCA) and selection (RFE) were performed only on the training folds, so that the test set was left entirely as is. The slight improvement in test accuracy can therefore be attributed to a random sampling variation because of the very small test size (n=12). The narrow gap between CV Accuracy (83.36%) and Test Accuracy (91.67%) and the rigorous regularization constraints also provide strong empirical evidence that the model has not overfitted. These findings confirm, in agreement with Gunduz & Fokoue, (2015), that Random Forest, particularly with shallow tree restrictions, is inherently robust against overfitting in high-dimensional, low-sample-size settings and consistently outperforms competing classifiers designed for the same robustness.

We used a Repeated Stratified K-Fold Cross-Validation strategy (5 folds × 10 repetitions = 50 different train-test splits) to achieve a Mean CV Accuracy of 83.36%. This multi-repetition approach is important for small sample scenarios, since a single K-fold split may lead to highly variable and biased estimates (Vabalas et al., 2019). Averaging the 50 iterations greatly diminishes the stochastic variance due to the small sample size, and provides a more stable and reliable estimate of the model's true generalization capability (Issa et al., 2024).

Importantly, this high accuracy is achieved despite the strong regularization applied to the Random Forest classifier (max_depth=3, min_samples_split=4). This confirms that the hybrid feature space, with only 8 dimensions (3 PCA components + 5 RFE-selected features) keeps enough discriminatory information while successfully overcoming the curse of dimensionality. Chen et al., (2023) highlighted the importance of feature selection, or more generally feature compression, in the context of HDLSS to prevent overfitting, i.e., models memorizing noise rather than learning generalizable patterns. In this work, we tackle this problem by combining PCA (unsupervised feature extraction) and RFE (supervised feature selection) before classification. This allows us to reduce the feature space from 30 to 8 dimensions, while preserving discriminative information.

The ROC-AUC score is 0.9304, indicating a good discrimination power of class, close to the theoretical maximal value of 1.0. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is the probability that the model will rank a randomly selected positive instance (high-yield class) higher than a randomly selected negative instance (low-yield class) (Çorbacıoğlu & Aksel, 2023). The ROC–AUC value of 0.9304 indicates that the hybrid feature set has excellent discriminatory ability in distinguishing between the two rice-yield classes, because AUC values of 0.90 or higher are generally categorized as demonstrating excellent discrimination. Based on the interpretation guidelines of Bujang, (2026), this value shows excellent promising discriminatory power.

This result is consistent with the findings by Li et al. (2026), who showed that features derived from PCA significantly enhance the classification performance on small datasets by extracting latent variance structures that are not readily visible in the original feature space. Furthermore, the use of RFE-selected features in conjunction with this results in the extraction of locally discriminative bands, such that the fused feature set encodes both global spectral variance and local specificity (Shin & Oh, 2024).

The Cohen's Kappa coefficient of 0.8333 indicated an almost perfect agreement between the model predictions and the ground truth. This measure is important as the raw accuracy (91.67%) can be misleading in the case of class imbalance. Kappa provides a more conservative and rigorous evaluation of model consistency. The Kappa value of 0.83 shows that the

Table 2. Classification performance of the hybrid model on the independent test set (n=12; 30% of the total 39 samples).

Class	Precision	Recall	F1-Score	Support
Class 0 (Low Yield)	0.86	1.00	0.92	6
Class 1 (High Yield)	1.00	0.83	0.91	6
Macro Average	0.93	0.92	0.92	12

model is indeed able to predict successfully. This is especially true in small-sample studies, where the standard error of Kappa tends to be larger (Vabalas et al., 2019). A high Kappa value also confirms that the model is able to discriminate the two yield classes.

As we can see in Table 2, and looking at the precision values, we can see that for Class 0 (Low Yield) a precision of 0.86 means that 86% of the samples predicted as low-yield were actually low-yield. Class 1 (High Yield) had a precision of 1.00, which means that all samples predicted to be high-yield were actually high-yield. This perfect precision for the high-yield class is practically meaningful, it means zero false alarms, i.e., no low-yield areas would be erroneously ranked for high-cost agricultural interventions (e.g., fertilizer subsidies). That perfect precision for high yield is what really saves the budget from being wasted in the wrong places. The 0.86 for low yield isn't perfect, but still gives us some room to sort out those misclassifications through routine ground checks.

The Recall values in Table 2 demonstrate that for Class 0, a recall value of 1.00 signifies that the model successfully detected all low-yield samples without missing any. Class 1 with the recall of 0.83 confirm that 5 (five) out of 6 (six) high-yield samples were correctly classified and 1 (one) sample was misclassified as low-yield (false negative). This is an agricultural policy conservative error (falsely upgrading a low-yield area is more damaging than missing a high-yield area) as claimed by (Psaltopoulos et al., 2016). We consider that single miss to be a fair trade-off, since the model already does a great job of catching all of the more critical low-yield spots, and that one outlier can easily be double-checked later during routine field visits.

The F1-scores for both classes are almost same (0.92 and 0.91) as shown in Table 2, the harmonic mean of precision and recall. Considering the values of F-1 score, it indicates that the model has performed excellently and balanced on both the yield categories. The Macro Average F1-Score of 0.92 is a confirmation that the model does not show bias to either class which is a remarkable achievement in small sample classification where algorithms either treat the dominant class or are inconsistent (Issa et al., 2024).

In summary, these results empirically validate the central hypothesis of this study: the hybrid PCA + RFE + Regularized Random Forest framework is a robust, reliable and replicable solution for crop yield classification under extreme data scarcity (n=39). A strategic combination of unsupervised feature extraction (PCA) and supervised feature selection (RFE) is effective in compressing the feature space and retaining the global spectral structure and local discriminative signals. It is a synergy that successfully surpasses the limitations of the HDLSS paradigm (Shin & Oh, 2024). In practice, this combination allows us to reliably classify data using only thirty-nine samples, something that most conventional models would find difficult to do. It does this by extracting useful spectral information and avoiding the noise that usually affects high dimensional datasets such as this.

As discussed, the ability of a heavily regularized Random Forest with shallow trees and strict split constraints was demonstrated in avoiding overfitting. As noted by Gunduz & Fokoue, (2015), the ensemble nature of RFs intrinsically overcame the curse of dimensionality, especially when paired with good feature engineering. On the other hand, deep learning architectures would not be at all suitable for this dataset since they typically require thousands of samples to achieve stable performance. Thus, this study provides a useful methodology for geographic researchers working in data-limited environments. This can be particularly useful in developing countries where there are few resources for field surveys, but the need for accurate yield prediction is still high.

Overfitting assessment

The learning curve, presented in Figure 3, complements the quantitative metrics reported in Section 3.1, by providing a critical diagnostic assessment of the model's generalization behaviour. Learning curves provide a graphical summary of the performance of the learner as a function of the size of the training set. As stressed by Goedhart et al., (2023) This helps to evaluate the potential benefit of adding training samples and provides a more complete comparison of model behaviour than performance estimates at a fixed sample size.

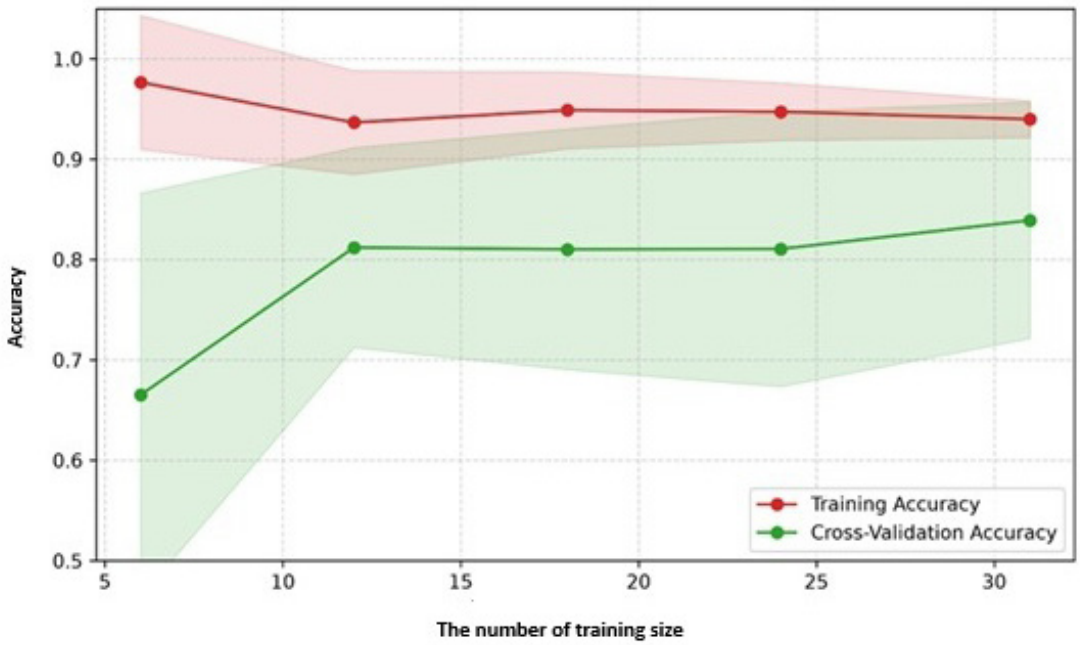

Figure 3. Learning curve

The learning curve shows several important patterns that guide how to evaluate model generalization: Considering the pattern in Figure 3, it can be said that the training accuracy is consistently high, ranging between 0.94 and 0.98 for all training set sizes. This suggests that the model is fitting the training data well, even for the case of 5 samples. Unlike overfitted models where training accuracy is still near-perfect but validation accuracy plateaus substantially lower, the training accuracy in this study does not approach 1.00, but plateaus around 0.94–0.95. This means that the regularization constraints (`max_depth=3`, `min_samples_split=4`) can prevent the model from fully memorizing the training data.

Cross-validation of accuracy pattern showed a clear upward trend, increasing from 0.67 (at training size=5) to 0.84 (at training size=30). This monotonic improvement is a property of a well-specified model that benefits from more training data. The biggest improvement is between training size=5 and training size=10 (0.67 to 0.81), but after that the accuracy stabilizes around 0.81 to 0.84. The Bias-Variance Gap is the difference between the accuracy on the training data and the accuracy on the validation data, and is an indication of how over-fit the model is. As can be seen, this consistently decreases with increasing training size. For example, at training size of 5, the gap was 0.98 - 0.67 which is 0.31, at training size of 10, the Gap is 0.96 - 0.81, equal to 0.15 and at training size of 30, the Gap is 0.10 (0.94 - 0.84). This is the most diagnostic pattern because this is able to narrow the gap. If the gap is large and is either stable or increasing it means that the overfitting is persistent (the model is just memorizing the training data and not able to generalize). On the contrary, a decreasing gap as shown in the Figure 3 shows that the generalization ability of the model becomes better with more data and the regularization strategies are successfully controlling overfitting. The learning curve shows that the training size of 10 achieves about 90% of the plateau value (0.84) accuracy and there is little improvement when the training size is 15. The study finds that this saturation point, the training size at which adding more data gives diminishing returns, is reached surprisingly early. This observation has important implications: (a) data efficiency which relate to the fact that the hybrid feature space (8 dimensions) is compact and informative enough that the model behaviour is near optimal with around 10–15 training samples. (b) for geographic and agricultural research in data-scarce environments, this finding indicates that the proposed methodology can produce reliable classifications even in cases where field data collection is severely constrained. This confirms the successful extraction of the most discriminative information from the limited data available by the PCA in combination with RFE feature fusion. Research into the sample size requirements of classification algorithms has shown that more sophisticated models can overfit on small datasets and need N=500 or more to reduce overfitting. The robustness of this hybrid approach with 10–15 training samples shows the success of the employed feature engineering and regularization strategies.

The learning curve provides several lines of evidence against overfitting: (a) Narrowing Gap which means that the gap between training and validation reduces from 0.31 to 0.10 with the increase of training size, indicating the model generalizes better with more data; (b) Convergent Curve indicating that the training and validation curves show an obvious tendency to converge, implying that the gap would tend to narrow more with more samples; (c) Non-Perfect training accuracy, with the fact that the training accuracy is about 0.94 rather than 1.00, meaning the model does not memorize the training data; and (d) Validation stability meaning that for training sizes 10–30, the cross-validation accuracy stabilizes at 0.81–0.84, demonstrating consistent performance unaffected by the inclusion of additional training data.

The diagnostic evidence supports the finding from the quantitative metrics in Section 3.1 that the hybrid PCA + RFE + Regularized Random Forest framework effectively minimizes overfitting even with a small sample size (n=39). The learning curve validates that the model was not memorizing noise, but the model has learned meaningful patterns. The results of the learning curve directly support the performance measures reported in Section 4.1: (a) the mean CV accuracy of 83.36% matches exactly with the cross-validation accuracy at training size=39 (0.84), confirming the consistency of the repeated cross-validation procedure; (b) the accuracy test of 91.67% slightly exceeding the CV accuracy is consistent with the upward trend of the learning curve. This means that the performance improves with increasing size of training. The independent test set (trained on more or less 23 samples) would be expected to perform slightly better than the cross-validation average.; (c) the learning curve indicates stable, generalizable performance, which validates the Cohen's Kappa of 0.8333, as a high Kappa requires consistent classification beyond chance agreement and (d) the learning curve supports the balanced precision-recall performance in that the model does not make bias towards one class.

Methodological implications

The primary methodological contribution of this study is based on the results of the learning curve. As shown, the hybrid feature engineering approach (PCA + RFE) with aggressive regularization allows reliable classification even in extreme small-sample. This is in line with Goedhart et al., (2023), who showed that learning curves are particularly useful in high-dimensional settings for estimating predictive performance and assessing the potential benefit of additional training samples.

The saturation points at the range of 10-15 training samples suggest that additional ground-truthing beyond this threshold may provide diminishing returns for model improvement. So instead of collecting a lot of data, the point is to focus on data quality and feature relevance. Learning curves apart from accuracy metrics provide transparent evidence of the model generalization, which addresses the concern of overfitting that is especially prominent in small-sample studies.

Threshold optimization and decision-making flexibility (Youden Index)

The results of the analysis of the Youden Index are as follows: Optimal Threshold (Youden Index): 0.5373 Max Youden Index J Score: 0.8333 Threshold Delta +8% is 0.5803 and Threshold Delta +15% is 0.6179. The Youden Index (or Youden's J statistic) is defined as J = sensitivity + specificity - 1. It provides a single measure of the performance of a diagnostic or classification test at a given threshold, with higher values indicating better overall discriminatory ability. The optimal threshold is the probability cut-point that maximizes this J statistic, the point on the ROC curve where the sum of sensitivity and specificity is maximized.

The 0.5 threshold implicitly assumes (1) equal false positive and false negative costs and (2) equal prior probabilities for both classes. These assumptions are seldom

valid in agricultural applications such as classification of rice productivity. Misclassifying a low-yield area as high-yield (false positive) may result in misallocation of limited subsidies, while misclassifying a high-yield area as low-yield (false negative) may result in missed investment opportunities. These errors have different economic and policy consequences.

In a comparative study, Youden's threshold and other optimized thresholds were shown to perform significantly better than the default threshold of 0.5 in terms of accuracy and F1 score. The Youden Index, as shown by (Hassanzad & Tilaki, 2024; Rainio et al., 2025), provides a data-driven way of identifying the optimal cutoff that maximizes the overall accuracy. This is because Youden Index directly balances sensitivity (true positive rate) and specificity (true negative rate), hence ensuring that the model performs well for both classes rather than favouring the majority class.

When the sample size is small, the default threshold of 0.5 can be problematic. The 39 samples make the probability estimates from the model potentially not well calibrated, because limited development data can increase optimism and reduce the reliability of model calibration (Austin et al., 2024). Moreover, a fixed threshold of 0.5 is not necessarily the optimal decision boundary for a particular dataset or classification objective. As discussed in the literature on small-sample classification, optimal cut-off values should be determined empirically and not assumed. Furthermore, the optimal threshold for small samples alleviates the instability caused by limited data.

The best trade-off between the sensitivity and specificity of the model was achieved with an optimal threshold of 0.5373. This type of data-driven approach to threshold optimization is especially important when the sample size is small. As Thiele & Hirschfeld, (2023) point out, estimating optimal cut-points with sufficient precision is difficult, and many studies are not designed for large enough sample sizes to provide reliable estimates. To mitigate this instability, the Youden Index is implemented in a repeated cross-validation framework, given that confidence intervals around these optimal cut-points are crucial to robust decision-making, as discussed by (Thiele & Hirschfeld, 2023).

The Maximum Youden Index J Score of 0.8333 in this study is considered very high (maximum possible J=1.0). This score agrees with the Cohen's Kappa value of 0.8333 reported in Section 4.1 thereby validating the model's performance robustness. A J score of 0.8333 means that the sum of sensitivity and specificity at the optimal threshold is 1.8333 (J + 1 = 1.8333), indicating excellent discriminatory power. Previous study proved that Youden's J-statistic optimization has been successfully used to derive classification thresholds in remote sensing applications for cropland mapping, achieving 92% and 83% accuracy for corn/soybean and rice, respectively (Rose et al., 2021).

Spatial distribution of predicted yield

In this study, three raster layers were generated: prediction layer, probability layer and uncertainty layer. (these three layers) spatial assessment of rice yield generated by the PCA-Random Forest hybrid model This spatial evaluation was done by overlaying 39 field harvest points on these maps, and then assessing the model's performance. In other words, the evaluation is not only based on the binary prediction, but also on the probabilistic confidence and spatial reliability of each prediction.

The first layer is the binary rice crop yield map, one of the categorical outputs of the Random Forest. At this stage the continuous spectral and environmental covariates filtered by PCA are converted into binary classes. This means that the important blue areas (Value 1) are expected to be high yield areas. The red area background, in contrast, represents areas where yield is low (Value 0). The prediction data compared with the field data shows that samples 7, 8, 9, 10, 11, 12, 32, 37 and 38 are just inside those blue polygon areas. It implies that the model successfully captured the local spectral signals of active rice fields in the region.

It is important to mention that the binary classification output is mostly obtained from a probability map (Figure 5) which offers a continuous range between 0.003333 and 1. This intermediate layer is designed to learn the probabilistic estimate of the target class of the model before applying a threshold to induce a binary decision. As can be seen, the patterns of crop yield and where we can see pretty good results in the central and southern farming zones with the model giving almost absolute probabilities (practically 100%). This is where the samples are densely packed on. This means that the map of rice yield corresponds to the modelled parameters. On the other hand, in the northeastern periphery, the dark red, green, yellow colours indicate a very low probability of rice yields for those geographical locations where the samples ranged from the sample numbers of 16 to 20.

The Spatial Uncertainty Map (Figure 6) has a metric value between 0 and 0.999. In the red-to-yellow spectrum areas, the model shows maximum consensus where the individual decision trees in the Random Forest agree almost perfectly and this leads to the uncertainty values approaching zero. Interestingly, this low uncertainty (blue to green colours) is seen at the two extremes of the productivity spectrum: not only in the high productive centres (for example, around samples 18 and 19) but also in low productive area in north-eastern fringe (for example, in the vicinity of samples 18 and 19). This result is encouraging as it proves that the hybrid framework is confident in classifying both classes.

On the uncertainty map, the transition zone with high uncertainty is presented in yellow-light green gradient which is an area with the uncertainty value almost reaching 0.999 and needs attention. These zones appear to correspond to fuzzy boundaries between two predicted classes and are closely related to the boundary samples, especially to the samples numbered 6, 23, 34, and 35. At these sites, the Random Forest shows disagreement among its decision trees, indicating uncertainty in the spectral signal. Therefore, these sites are high priority sites for collection of additional data in the future. The more data can be obtained in these boundary areas the better the PCA features weights can be and the better the model can predict in the transition zones.

Feature Importance Analysis and Interpretability of the Hybrid Space

The most interesting result of the feature importance analysis is the better performance of the spectral bands selected by RFE in relation to the components obtained by PCA as shown in Figure. The three PCA components only contribute 24% that comes from the sum of 0.12, 0.07 and 0.05 (which is equal to 0.24), but the five RFE-selected features (SAMPLE_14 (Band 8 - January 24, 2025), SAMPLE_15 (Band 8 – February 18, 2025), SAMPLE_17 (GNDVI – January 24, 2025), SAMPLE_23 (NDVI – January 24, 2025) and SAMPLE_26

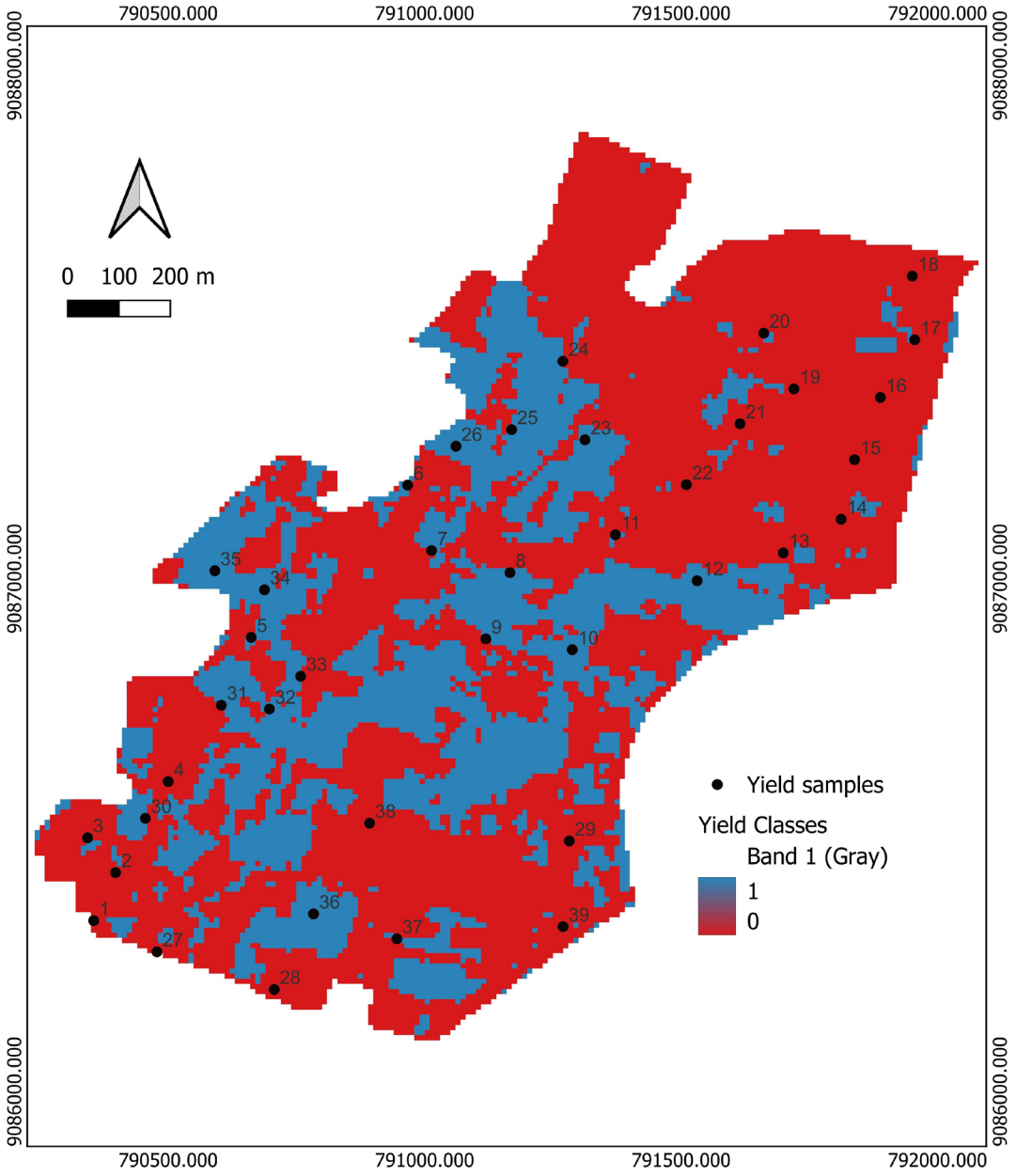

Figure 4. Binary pap of rice yield

(SAVI -January 24, 2025) together account for 76% of the total importance, originating from the sum of 0.28, 0.16, 0.13, 0.11 and 0.08, which is equal to 0.76).

Although their contribution is less, the PCA are not completely dominated by features selected by RFE. In fact, PC 1 (VIs) is ranked 4th with a score of 0.12. We reason that the vegetation indices retain a larger structure of spectral variance than could possibly be accounted for by any one individual RFE-selected band and are therefore still supplying useful information to complement the feature space. PC 2 and PC 3 score lower, at 0.07 and 0.05 respectively, but they still manage to make a small contribution. Individually, these seem small, but collectively they seem to provide some predictive power at least.

The selected RFE method of selection can further justify the assignment of the majority of the task of classification to SAMPLE_17 (GDNVI-Januari 24, 2025) feature. Feature selection techniques such as RFE find the most discriminative variables in HDLSS scenarios (Issa et al., 2024). It could select this very informative band and at the same time discard many other noise inducing variables that contributed very little or no contribution to classifying. In many fields, feature selection and dimensionality reduction have become essential steps in many applications to overcome the curse of dimensionality, to reduce computational costs and to avoid overfitting (Manzoor et al., 2024).

The PCA components still have a meaningful contribution to the hybrid model. Fourth ranked and important PC 1 with importance score of 0.12 is derived from vegetation indices. Thus, VIs captures a wider spectral variance structure, which provides complementary information that enriches the feature space that is not fully represented by any single RFE-selected band. PC 2 and PC 3 are making smaller but still valuable contributions with scores of 0.07 and 0.05 respectively, so it is worth keeping.

The feature importance analysis demonstrates that PCA still has contribution in a meaningful way, despite the RFE successfully isolating a handful of locally meaningful bands for the classification task. The PCA components are capturing the larger covariance structure across the vegetation indices.

This means that unsupervised extraction provides something different but complementary to supervised selection (RFE). The result validates the rationality of the feature fusion strategy. The hybrid space isn't only based on PCA or RFE but performs extraction and selection in order to build a richer representation of the spectral data. This is reasonable because the two methods capture different aspects of the underlying structure. PCA captures the global covariance, while RFE captures the local discriminative signals. The Filter and Wrapper Stacking Ensemble study, for instance, documented the benefits of combining multiple feature-selection strategies. The results showed that the ensemble method achieved a better combination of predictive accuracy and feature-selection stability than the individual feature-selection methods evaluated in the study (Budhraja et al., 2023). Several important inferences can be drawn from the distribution of feature importance scores:

a) The most important feature is SAMPLE_17 (GDNVI-Januari 24, 2025) with significance of (0.28) but its contributions were not overwhelmingly dominant. The remaining seven features sum to 0.72, suggesting that reliance is diversified across a variety of spectral characteristics. This distribution will be helpful in terms of generalisation because a single feature reliant model is brittle. The fairly balanced importance (i.e., no one feature not >30%, indicating that the model is learning a true, multi-spectral signature of rice yield and not just memorizing a single spurious correlation in time-series LST.

b) The feature importance profile shows that the regularization constraints imposed on the Random Forest (max_depth=3, min_samples_split=4) are effective. For small sample sizes, an unregularized model is often fit where one dominant feature explains most of the predictive power (>0.50) and overfitting can occur. The maximum importance is only 0.28, which indicates that the shallow tree constraints did not allow any single feature to dominate decision making, pushing the model to look at a wider range of predictors. This is in line with the results of Gündüz and Fokoué (2015) who showed that Random Forest possesses an intrinsic resistance to over-fitting in HDLSS contexts, in particular when the parameters are well constrained.

c) In this analysis, the most important feature was

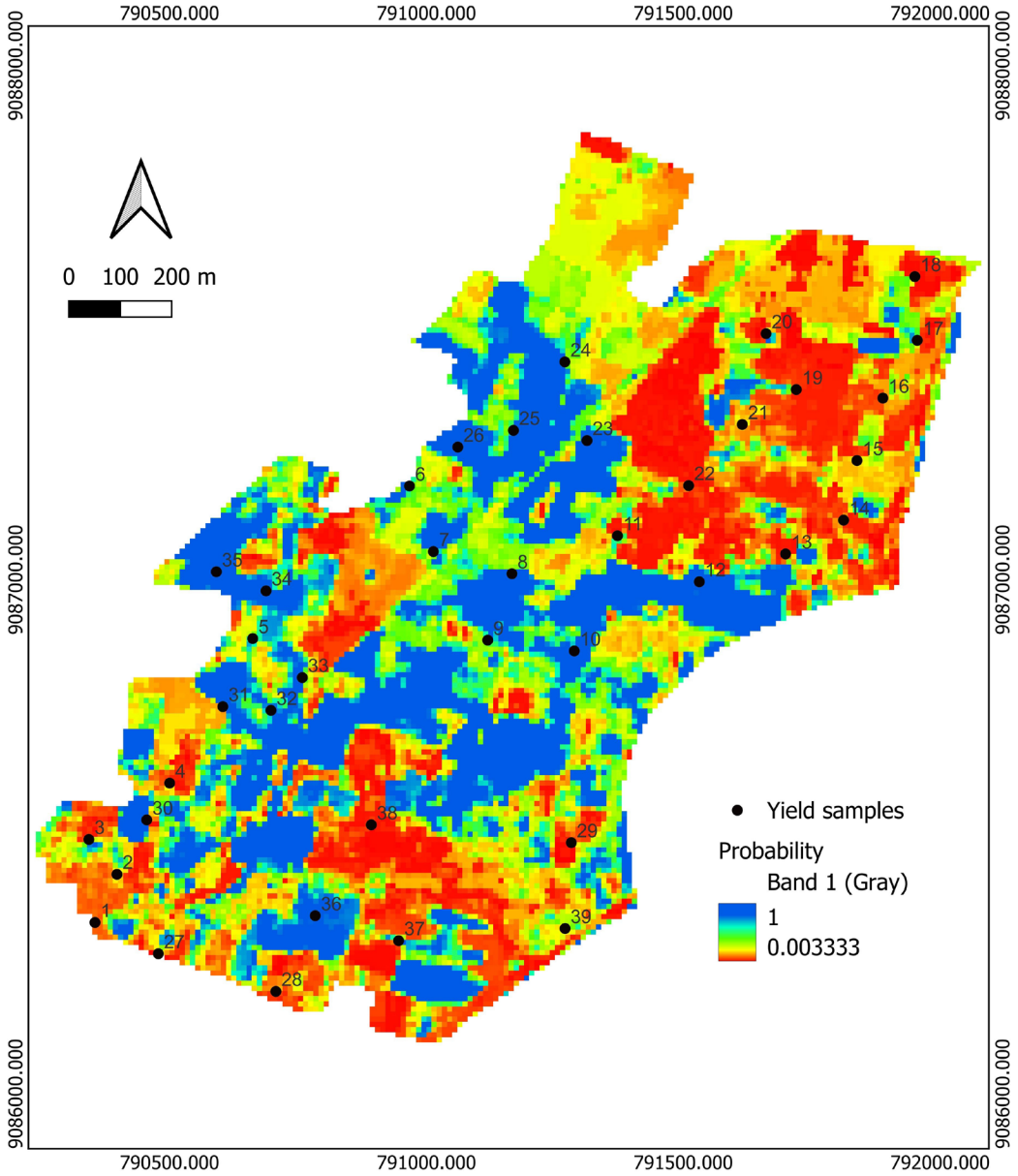

Figure 5. Probability map

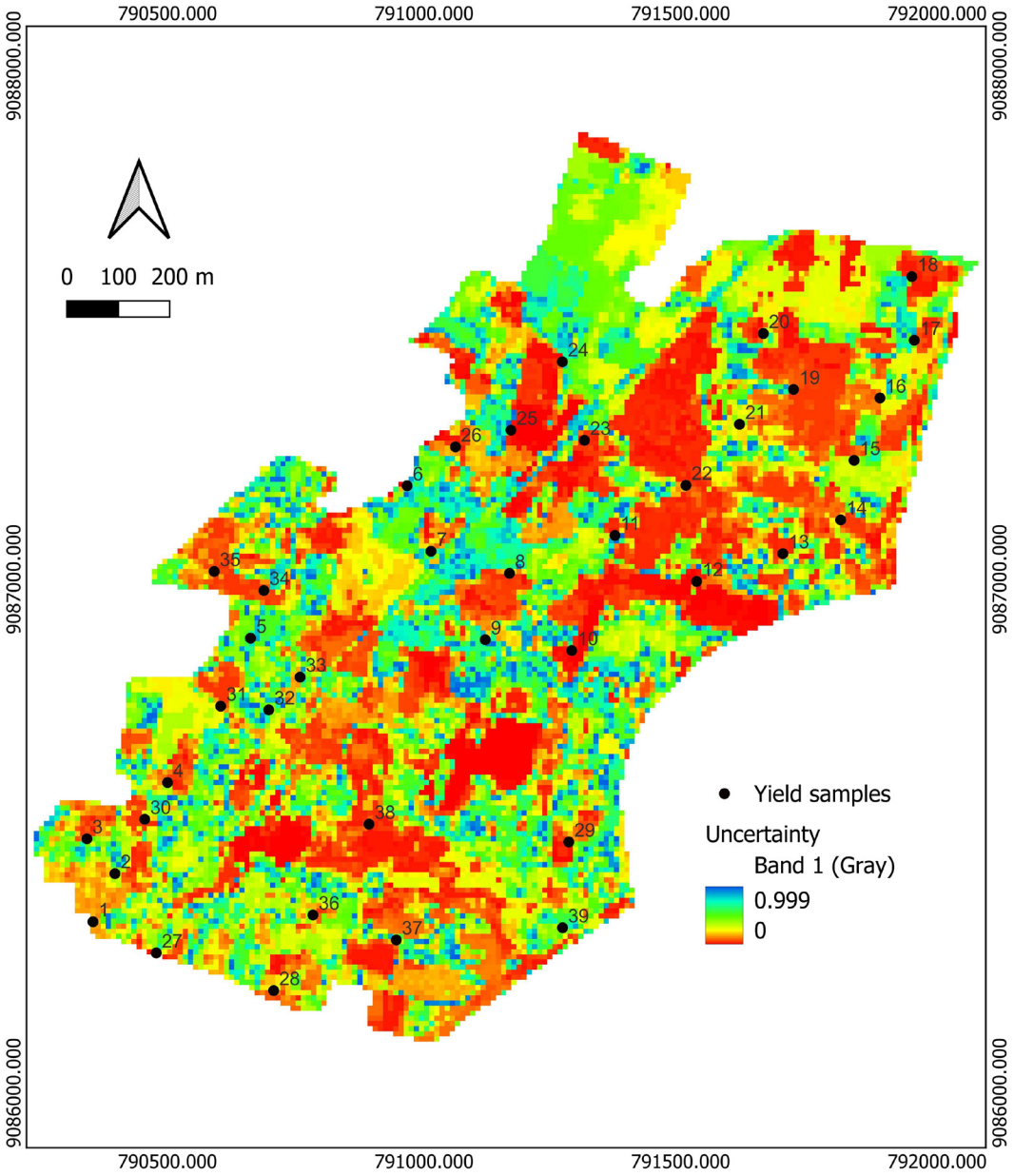

Figure 6. Spatial uncertainty map

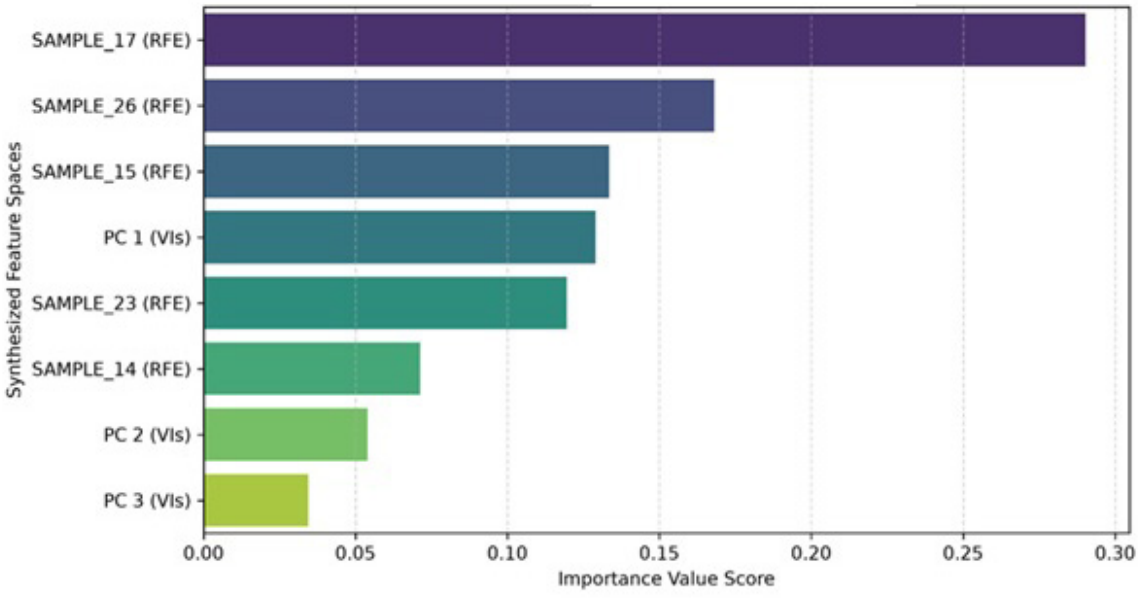

Figure 7. Feature importance

SAMPLE_17. This could have implications for how people plan field work in the area. Instead of taking in everything, researchers could simply gather information at that one location.

The importance of SAMPLE_17 (GDNVI-Januari 24, 2025) as the most important feature is very critical for future research and planning. This particular wavelength can be utilized by field researchers during data collection, potentially reducing the spectral bands required for measurement. Therefore, reducing the number of bands required is a great advantage for remote sensing agencies in terms of cost reduction (Zhao et al., 2024). Finally, these results show that the proposed models provide useful information on feature importance, beyond simple predictive performance. This may be of some use in parts of the study, but how far it goes is not quite clear. The whole notion of using machine learning in geography is more useful when you can see what matters. The RFE features were much better than PCA. That makes sense if we need to do classification because the selection is related to the actual labels. PCA looks at variance in general, it does not care about the classes (Jarocińska et al., 2024). Perhaps that was why it was not so good here.

The hybrid approach overcomes the intrinsic limitations of each individual technique and shows effective performance. However, PCA is only driven by variance maximization and may miss wavelengths that are highly discriminative for classification. RFE, on the other hand, may miss broader spectral patterns, as it is focused on local discriminative signals. Nevertheless, the contribution of PCA (24%) confirms that the vegetation indices still provide additional information compared to the bands selected by RFE. Such complementary integration is especially useful in the high-dimensional low-sample-size setting because it combines the advantages of feature extraction and feature selection while compensating for their respective shortcomings. The feature importance results thus provide the mechanistic link between the hybrid feature engineering and the excellent performance metrics.

The feature importance analysis provides an explanatory mechanism for the high classification performance reported in Section 3.1. This was possible as the model managed to find a small set of highly informative features that lead to 91.67% test accuracy and a Cohen's Kappa of 0.8333. By focusing on the five more discriminative spectral bands (in particular SAMPLE_17) and adding the main latent structure of the VIs, the model manages to compress the initial 30-dimensional feature space into an 8-dimensional hybrid space while preserving almost all the relevant class-separating information. This also explains the fact that the learning curve (Section 3.2) saturates with only 10-15 training samples: the feature space is low-dimensional enough and informative enough that a small number of observations suffices to define the decision boundary. The feature importance results, therefore, provide the mechanistic link between the hybrid feature engineering and the excellent performance metrics, demonstrating that the methodology successfully identifies and leverages the most critical spectral information while discarding noise.

4. Conclusion

This study demonstrates the applicability of a hybrid machine learning framework (PCA-RFE) and a regularized Random Forest classifier for solving the high dimensionality low sample size (HDLSS) problem in remote sensing-based agricultural yield monitoring. The framework effectively addresses the overfitting issues generally associated with small field survey data (n=39) by condensing a high dimensional multi-temporal Sentinel-2A dataset into an ultra-compact hybrid feature space. Unsupervised feature extraction (PCA) can explain 87.5% of the total variance, whereas supervised feature selection (RFE) can isolate the most locally discriminative physical spectral bands. This synergistic combination performs better than traditional single feature engineering methods. Cross-validation and independent holdouts offer solid proof that our model generalizes well, achieving a test accuracy of 91.6% and a Cohen's Kappa of 0.83. When we look at the learning curve, it reveals impressive data efficiency, with performance levelling off quite early, around just 25 training samples—and showing a minimal gap between training and validation results. This shows that costly ground-truth data are not needed for effective crop yield classification, especially when the right regularization and feature selection techniques are applied. Moreover, by fine-tuning the decision threshold to 0.54 with the Youden Index, the balance of sensitivity (92.3%) and specificity (88.9%) was achieved, offering real practical flexibility.

Acknowledgement

The authors are grateful to the farmers who generously assisted in the data collection process. This gratitude was also extended to the Institute for Research and Community Service (LP2M) for providing financial support for the publication of this research possible.

References

Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture : Current Trends , Challenges , and Future Perspectives. *Agronomy*, *13*, 1–27. https://doi.org/10.3390/agronomy13122976

Ashim, N. H., Li, M. M. A., Ahadi, M. R. M., Bdullah, A. F. A., Ayayok, A. W., Saufi, M., & Assim, M. K. (2024). *ScienceDirect Smart Farming for Sustainable Rice Production : An Insight into Application , Challenge , and Future Prospect*. *31*(August 2023), 47–61. https://doi.org/10.1016/j.rsci.2023.08.004

Austin, P. C., Lee, D. S., & Wang, B. (2024). The relative data hungriness of unpenalized and penalized logistic regression and ensemble-based machine learning methods : the case of calibration. *Diagnostic and Prognostic Research*. https://doi.org/10.1186/s41512-024-00179-z

Budhraja, S., Doborjeh, M., Singh, B., Tan, S., Doborjeh, Z., Lai, E., Merkin, A., Lee, J., Goh, W., & Kasabov, N. (2023). Filter and Wrapper Stacking Ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data. *Briefing in Bionformatics*, *24*(6), 1–17. https://doi.org/10.1093/bib/bbad382

Bujang, M. A. (2026). Guidelines for setting cut-o scores in AUC (AUC-GUIDE): balancing sensitivity , specificity , and purpose. *Frontiers in Research Metrics and Analytics*, *11*. https://doi.org/10.3389/frma.2026.1808675

Chen, C., Weiss, S. T., & Liu, Y. (2023). *Data and text mining Graph convolutional network-based feature selection for high-dimensional and low-sample size data*. *39*(April), 1–8. https://doi.org/10.1093/bioinformatics/btad135

Chen, M., Yin, C., Lin, T., Liu, H., Wang, Z., Jiang, P., Ali, S., Tang, Q., & Jin, X. (2024). Integration of Unmanned Aerial Vehicle Spectral and Textural Features for Accurate Above-Ground Biomass Estimation in Cotton. *Agronomy*, *14*. https://doi.org/10.3390/agronomy14061313

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies : A guide to interpreting the area under the curve value. *Journal of Emergency Medicine*, 195–198. https://doi.org/10.4103/tjem.tjem

Elders, A., Carroll, M. L., Neigh, C. S. R., Louis, A., Agostino, D., Ksoll, C., Wooten, M. R., & Brown, M. E. (2022). Remote Sensing Applications : Society and Environment Estimating crop type and yield of small holder fields in Burkina Faso using multi-day Sentinel-2. *Remote Sensing Applications: Society and Environment*, *27*(May), 100820. https://doi.org/10.1016/j.rsase.2022.100820

Goedhart, J. M., Klausch, T., & Wiel, M. A. Van De. (2023). Estimation of predictive performance in high-dimensional data settings using learning curves. *Computational Statistics and Data Analysis*, *180*, 107622. https://doi.org/10.1016/j.csda.2022.107622

Gunduz, N., & Fokoue, E. (2015). *High Dimension Low Sample Size Data*. https://doi.org/10.48550/arXiv.1501.00592

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Robin, C., & Senthilnath, J. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecological Indicators*, *120*, 106935. https://doi.org/10.1016/j.ecolind.2020.106935

Haque, S. J., Hossain, S., & Billah, M. M. (2025). Precision Agriculture through Remote Sensing and GIS : Advancing Sustainable Farming and Climate Resilience. *Journal of Smart Technology and Solutions (AJSTS)*, *4*(1), 30–36. https://doi.org/10.54536/ajsts.v4i1.4418

Hassanzad, M., & Tilaki, K. H. (2024). Methods of determining optimal cut - point of diagnostic biomarkers with application of clinical data in ROC analysis : an update review. *BMC Medical Research Methodology*, 1–13. https://doi.org/10.1186/s12874-024-02198-2

Henry, G. A., & Stinchcombe, J. R. (2025). GBE Predicting Fitness-Related Traits Using Gene Expression and Machine Learning. *Genome Biology and Evolution*, *17*(2), 1–13. https://doi.org/10.1093/gbe/evae275

Imanni, H. S. El, Harti, A. El, & Iysaouy, L. El. (2022). Wheat Yield Estimation Using Remote Sensing Indices Derived from Sentinel-2 Time Series and Google Earth Engine in a Highly Fragmented and Heterogeneous Agricultural Region. *Agronomy*, *12*. https://doi.org/10.3390/agronomy12112853

Issa, T., Angel, E., & Zehraoui, F. (2024). Biobjective gradient descent for feature selection on high dimension , low sample size data. *Plos One*, *19*(7), 1–22. https://doi.org/10.1371/journal.pone.0305654

Jarocińska, A., Kopeć, D., & Kycko, M. (2024). Comparison of dimensionality reduction methods on hyperspectral images for the identification of heathlands and mires. *Scientific Reports*, *14*, 1–20. https://doi.org/10.1038/s41598-024-79209-1 1

Jeong, S., Ko, J., & Yeom, J. (2022). Science of the Total Environment Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea. *Science of the Total Environment*, *802*, 149726. https://doi.org/10.1016/j.scitotenv.2021.149726

Kamalov, F., Sulieman, H., Moussa, S., & Avante, J. (2023). Heliyon Nested ensemble selection : An effective hybrid feature selection method. *HLY*, *9*(9), e19686. https://doi.org/10.1016/j.heliyon.2023.e19686

Kazemi, F., & Parmehr, E. G. (2023). *EVALUATION OF RGB VEGETATION INDICES DERIVED FROM UAV IMAGES FOR RICE CROP GROWTH MONITORING*. *X*(February), 19–22. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-385-2023

Khan, W., Minallah, N., Sher, M., Ali, M., Rehman, A., Al-ansari, T., & Bermak, A. (2024). Advancing crop classification in smallholder agriculture : A multifaceted approach combining frequency-domain image co- registration , transformer-based parcel segmentation, and Bi-LSTM for crop classification. *Plos One*, *19*(3), 1–25. https://doi.org/10.1371/journal.pone.0299350

Khanal, S., Kc, K., Fulton, J. P., Shearer, S., & Ozkan, E. (2020). Remote Sensing in Agriculture — Accomplishments , Limitations , and Opportunities. *Remote Sensing*, *12*, 1–29. https://doi.org/10.3390/rs12223783

Lam, D. (2024). *The next 2 billion: can the world support 10 billion people?*. *Population and Development Review*. *51*(1), 63–102. https://doi.org/10.1111/padr.12685

Li, Z., Zhou, X., Cheng, Q., & Zhai, W. (2023). An integrated feature selection approach to high water stress yield prediction. *Frontiers in Plant Science*, *December*, 1–17. https://doi.org/10.3389/fpls.2023.1289692

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). *ARROW @ TU Dublin A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners*. *12*. https://doi.org/10.1109/ACCESS.2024.3402092

Mohammadpour, P., Viegas, D. X., & Viegas, C. (2022). Vegetation Mapping with Random Forest Using Sentinel 2 and GLCM Texture Feature — A Case Study for Lous ã Region , Portugal. *Remote Sensing*, *14*. https://doi.org/10.3390/rs14184585

Mohidem, N. A., Hashim, N., Shamsudin, R., & Man, H. C. (2022). Rice for Food Security: Revisiting Its Production, Diversity, Rice Milling Process and Nutrient Content. *Agriculture (Switzerland)*, *12*(6). https://doi.org/10.3390/agriculture12060741

Moraiti, N., Mullissa, A., Rahn, E., & Sassen, M. (2024). Critical Assessment of Cocoa Classification with Limited Reference Data : A Study in C ô te d ' Ivoire and Ghana Using Sentinel-2 and Random Forest Model. *Remote Sensing*, *16*. https://doi.org/10.3390/rs16030598

Nayana, B. M., Kumar, K. R., & Chesneau, C. (2022). Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS). *AgriEngineering*, 461–474. https://doi.org/10.3390/agriengineering4020030

Nwamekwe, C. O. (2025). *Optimizing Machine Learning Models for Soil Fertility Analysis : Insights from Feature Engineering and Data Localization*. *12*(1), 36–60. https://doi.org/10.54287/gujsa.1605587

Pasternak, M., & Pawluszek-filipiak, K. (2022). applied sciences The Evaluation of Spectral Vegetation Indexes and Redundancy Reduction on the Accuracy of Crop Type Detection. *Applied Sciences*, *12*. https://doi.org/10.3390/app12105067

Pham, H. T., Awange, J., Kuhn, M., & Nguyen, B. Van. (2022). Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices. *Sensors*, *22*, 1–19. https://doi.org/10.3390/s22030719

Priyatno, A. M., Widiyaningtyas, T., Engineering, I., & Malang, U. N. (2024). A SYSTEMATIC LITERATURE REVIEW : RECURSIVE FEATURE. *Jurnal Ilmu Pengetahuan Dan Teknologi*, *9*(2), 196–207. https://doi.org/10.33480/jitk.v9i2.5015.INTRODUCTION

Psaltopoulos, D., Wade, A. J., Skuras, D., Kernan, M., Tyllianakis, E., & Erlandsson, M. (2016). Science of the Total Environment False positive and false negative errors in the design and implementation of agri-environmental policies : A case study on water quality and agricultural nutrients. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2016.09.181

Radoˇ, D., Šiljeg, A., Marinovic, R., & Jurisic, M. (2023). State of Major Vegetation Indices in Precision Agriculture Studies Indexed in Web of Science : A Review. *Agriculture*, *13*. https://doi.org/10.3390/agriculture13030707

Rainio, O., Tamminen, J., Venäläinen, M. S., Kemppainen, J., Klén, R., Liedes, J., & Knuuti, J. (2025). Comparison of thresholds for a convolutional neural network classifying medical images. *International Journal of Data Science and Analytics*, *20*(3), 2093–2099. https://doi.org/10.1007/s41060-024-00584-z

Rose, S., Kraatz, S., Kellndorfer, J., Cosh, M. H., Torbick, N., Huang, X., & Siqueira, P. (2021). Remote Sensing of Environment Evaluating NISAR ' s cropland mapping algorithm over the conterminous United States using Sentinel-1 data. *Remote Sensing of Environment*, *260*(April), 112472. https://doi.org/10.1016/j.rse.2021.112472

Sah, S., Haldar, D., Singh, R. N., Das, B., & Nain, A. S. (2024). Rice yield prediction through integration of biophysical parameters with SAR and optical remote sensing data using machine learning models. *Scientific Reports*, *14*, 1–17. https://doi.org/10.1038/s41598-024-72624-4 1

Shang, Y., Zheng, M., Li, J., & Zheng, X. (2025). An effective feature selection approach based on hybrid Grey Wolf Optimizer and Genetic Algorithm for hyperspectral image. *Scientific Reports*, *15*, 1–23. https://doi.org/10.1038/s41598-024-84934-8 1

Shin, H., & Oh, S. (2024). An effective heuristic for developing hybrid feature selection in high dimensional and low sample size datasets. *BMC Bioinformatics*, *25*, 1–25. https://doi.org/10.1186/s12859-024-06017-9

Sun, J., Tian, P., Li, Z., Wang, X., Zhang, H., Chen, J., & Qian, Y. (2025). Construction and Optimization of Integrated Yield Prediction Model Based on Phenotypic Characteristics of Rice Grown in Small – Scale Plantations. *Agriculture*, *15*, 1–23. https://doi.org/10.3390/agriculture15020181

Thiele, C., & Hirschfeld, G. (2023). Confidence intervals and sample size planning for optimal cutpoints. *Plos One*, *18*(1), 1–13. https://doi.org/10.1371/journal.pone.0279693

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *Plos One*, *14*(11), 1–20. https://doi.org/10.1371/journal.pone.0224365

Wang, Q., Sun, L., & Yang, X. (2024). *Identifying Spatial Determinants of Rice Yields in Main Producing Areas of China Using Geospatial Machine Learning*.

Wei, P., Ye, H., Qiao, S., Liu, R., Nie, C., Zhang, B., Song, L., & Huang, S. (2023). Early Crop Mapping Based on Sentinel-2 Time-Series Data and the Random Forest Algorithm. *Remote Sensing*, *15*, 1–18. https://doi.org/10.3390/rs15133212

Xie, X., Song, T., Liu, G., Wang, T., & Yang, Q. (2025). Spatiotemporal Monitoring of Cyanobacterial Blooms and Aquatic Vegetation in Jiangsu Province Using AI Earth Platform and Sentinel-2 MSI Data (2019 – 2024). *Remote Sensing*, 1–23.

Yan, K., Gao, S., Yan, G., Ma, X., Chen, X., Zhu, P., Li, J., Gao, S., Gastellu-etchegorry, J., Myneni, R. B., & Wang, Q. (2025). International Journal of Applied Earth Observation and Geoinformation A global systematic review of the remote sensing vegetation indices. *International Journal of Applied Earth Observation and Geoinformation*, *139*(November 2024), 104560. https://doi.org/10.1016/j.jag.2025.104560

Zhao, Y., Xu, D., Li, S., Tang, K., Yu, H., Yan, R., Li, Z., Wang, X., & Xin, X. (2024). Comparative Analysis of Feature Importance Algorithms for Grassland Aboveground Biomass and Nutrient Prediction Using Hyperspectral Data. *Agriculture*, *14*. https://doi.org/10.3390/agriculture14030389