Univerzitet u Beogradu
Poljoprivredni fakultet
Institut za poljoprivrednu tehniku
Naučni časopis
POLJOPRIVREDNA TEHNIKA
Godina L
Broj 3, 2025.
Strane: 1 – 13

University of Belgrade
Faculty of Agriculture
Institute of Agricultural Engineering
Scientific Journal
AGRICULTURAL ENGINEERING
Year L
No. 3, 2025.
pp. 1 - 13

Original Scientific paper
Originalni naučni rad
DOI:10.5937/POLJTEH2503001M

APPLICATION OF GLOBAL SATELLITE DATA AND MACHINE LEARNING MODELS FOR CROP CLASSIFICATION AND AGRO-ZONE IDENTIFICATION

Nataša S. Milosavljević*[1], Rade Radojević[1]

[1]University of Belgrade, Faculty of Agriculture, Institute of Agricultural Engineering, Belgrade-Zemun, Republic of Serbia

Abstarct: This study presents an applied approach to leveraging globally available satellite data—specifically the WorldCereal dataset—for the classification and spatial analysis of agricultural land. The dataset, developed under the European Space Agency (ESA), includes crop type, cropland probability, seasonality, and irrigation status for the year 2021, at 10-meter resolution. After filtering for high-confidence agricultural zones, we implemented Random Forest and XGBoost algorithms to classify crop types, and unsupervised K-means clustering to identify agro-zones with similar characteristics. Major crop categories included cereals, maize, rice, and pulses. The methodology was tested on a large subset of the global dataset, using a Python-based pipeline. Our results demonstrate that machine learning models can achieve high classification accuracy (F1 ≥ 0.98) even with limited input features, and that clustering techniques can reveal consistent spatial patterns without prior labeling. These findings highlight the potential of open-access satellite data and data-driven modeling in building cost-effective and scalable systems for precision agriculture, particularly in regions lacking detailed ground-truth information.

Keywords: *Remote sensing, land cover classification, tree-based learning algorithms, spatial data analysis, irrigation mapping, machine learning*

*Corresponding Author. Email address: natasam@agrif.bg.ac.rs
ORCID number: 0003-4056-089X
This research was supported by Ministry of Science, Technological development, and Innovations No. 451-03-137/2025-03/200116.
The grant agreement registration for the Faculty of Agriculture

INTRODUCTION

Agriculture today faces increasingly complex challenges, ranging from climate change and land degradation to the need for sustainable use of natural resources.

Within this context, numerous studies have positioned precision agriculture as a data-driven paradigm that integrates modern information and communication technologies (ICT) to enhance crop productivity, reduce environmental impact, and support sustainable resource management [11],[9],[8]. Key components of this paradigm include site-specific monitoring, variable-rate application of inputs, and real-time decision support systems. Remote sensing technologies, including satellite and UAV-based platforms, play a critical role by enabling timely and accurate mapping of field conditions at large spatial and temporal scales [9],[3],[4]. These capabilities provide essential inputs for optimizing seeding, fertilization, irrigation, and harvesting practices.

One notable initiative in this domain is WorldCereal, supported by the European Space Agency (ESA). This program aims to deliver global-scale datasets on crop types, cropland probability, seasonality, and irrigation by utilizing data from Sentinel-1 and Sentinel-2 missions. The resulting products are open-access, highly interoperable, and suitable for integration into diverse analytical systems [10].

Previous studies have shown that satellite data can successfully identify certain vegetation types. However, the application of machine learning algorithms to process publicly available global datasets is still emerging. Most existing research relies on regional or commercial data sources, while the use of open-source satellite layers like WorldCereal remains limited and primarily in the phase of validation and experimental deployment.

In this study, we propose the integration of the WorldCereal open dataset with both classical and advanced machine learning algorithms, including:

- Random Forest and XGBoost for crop type classification,
- K-means clustering for identifying agro-climatic patterns,
- and feature importance evaluation to support interpretability and decision-making.

Additionally, the study explores the potential for extending this methodology to other areas of precision agriculture, such as:

- detecting drought-prone regions,
- assessing erosion and land degradation risk,
- and fusing satellite data with drone imagery and IoT sensor networks at the local scale.

Our objective is to demonstrate how publicly available and globally standardized satellite data—such as the WorldCereal dataset—combined with open-source machine learning methods (e.g., Random Forest and XGBoost implemented in Python), can support the development of agricultural monitoring systems in data-scarce regions. This approach eliminates the need for expensive field equipment or proprietary software, making it suitable for use in countries with limited access to remote sensing infrastructure, local sensors, or high-performance computing resources.

The next section provides a detailed description of the materials and methods used in this study. Section materials and methods presents both quantitative and visual analyses of the classification and clustering results.

In the results and discussion section, we interpret the findings, examine limitations, and propose potential applications and integrations. Finally, Section conclusion summarizes the key insights and offers recommendations for future research directions.

MATERIALS AND METHODS

In this study, we conducted an empirical investigation of global cropland classification and agro-zone identification using the WorldCereal: Global Crop Type Product dataset at 10m resolution. This dataset, developed by the European Space Agency (ESA) and VITO Remote Sensing, served as the primary input for training supervised machine learning models (Random Forest, XGBoost) and applying unsupervised clustering techniques (K-means). Our research focuses on evaluating the performance of these models in accurately predicting crop types and delineating homogeneous agricultural zones using only satellite-derived attributes [1][2][4].

Key characteristics of the dataset include:

- Spatial resolution: 10 meters
- Geographic coverage: Global, with data available for agricultural regions worldwide
- Temporal coverage: Year 2021 (current version)
- Included layers (attributes):
- cropland_probability – probability that the area is cropland
- crop_type – crop classification (e.g., cereals, rice, maize)
- season – cropping season (spring, summer, winter)
- irrigated – irrigation status
- lat, lon – geographical coordinates
- confidence – classification confidence score
- additional NDVI layers derived from Sentinel-2 imagery (externally computed)

The data were downloaded in CSV format and prepared for analysis using Pandas and GeoPandas libraries in Python. For spatial visualization, folium, matplotlib, and seaborn were employed.

Upon acquisition, the dataset underwent a comprehensive preprocessing pipeline to ensure suitability for analytical modeling and evaluation. The first step involved removal of missing values in critical columns such as crop_type and cropland_probability, as these served as the foundation for both classification and clustering tasks.

Since the dataset comprised both numerical and categorical features (e.g., season, irrigated), categorical attributes were encoded into numeric representations, enabling seamless integration into machine learning pipelines. Particular attention was given to seasonal attributes, which carry strong spatial-temporal signals relevant to crop occurrence and dynamics [8].

Simultaneously, spatial distribution patterns were examined. Geographical coordinates (latitude and longitude) were normalized to eliminate outliers, and the data were spatially filtered to retain only agricultural zones with a high cropland probability, in alignment with the cropland_probability layer.

Based on externally derived NDVI values (Normalized Difference Vegetation Index) from Sentinel-2, additional variables were constructed to quantify vegetative intensity [3]. These features were instrumental in the unsupervised clustering phase aimed at identifying agro-ecological zones with similar characteristics [3].

To further refine class separability, binary indicators for irrigation status and seasonal cropping patterns were generated and incorporated into the modeling process [4].

The entire dataset was split into training and testing sets (70:30 ratio) while preserving the distribution of crop types across both subsets. This ensured representative sampling and reduced the risk of model bias. Standard preprocessing techniques such as scaling and label encoding were applied in accordance with best practices for implementing Random Forest and XGBoost classifiers.

This comprehensive data preparation laid a robust foundation for the deployment of classification models and enabled additional unsupervised learning via K-means clustering, which uncovered latent spatial patterns related to cropland distribution and crop diversity.

Random Forest is an ensemble learning method that builds a collection of decision trees during training and outputs the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. The technique increases robustness and reduces the risk of overfitting by introducing two levels of randomness: sampling of the training data (bootstrapping) and random feature selection for each split in a tree.

For a given input instance $x \in R^d$, the prediction $\hat{y}$ is obtained as:

$$\hat{y} = majority\_vote\{h_1(x), h_2(x), ..., h_T(x)\} \quad \text{.....(1)}$$

where $h_t(x)$ denotes the prediction of the t-th decision tree in the ensemble, and T s the total number of trees (in this research is T =50). Each tree is trained on a bootstrapped sample $D_t \subset D$, and at each node split, a random subset of features $F_t \subset F$ is considered, where:

$$|F_t| = \sqrt{|F|} \quad \text{.....(2)}$$

for classification problems (common heuristic), ensuring diversity among the trees.

The model also provides feature importance estimates based on the average decrease in impurity (e.g., Gini index) across all trees, which is especially useful in the interpretation of high-dimensional datasets, such as satellite-derived agricultural layers.

XGBoost (eXtreme Gradient Boosting) is a high-performance implementation of gradient boosting decision trees, designed for speed, scalability, and predictive accuracy. The term "eXtreme" reflects several key innovations that go beyond classical gradient boosting:

- incorporation of regularization terms (L1 and L2) to reduce overfitting,
- parallelized tree construction,
- sparse-aware handling of missing data, and
- optimized memory usage and support for distributed computing environments [Wikipedia, NVIDIA].

Unlike traditional boosting methods that sequentially build weak learners to minimize prediction error, XGBoost employs a second-order Taylor expansion of the loss function, which allows it to incorporate both gradients and Hessians into the optimization process. This yields more precise model updates and significantly improves convergence speed.

The objective function in XGBoost balances model fit with complexity control by adding a regularization term that penalizes overly deep or complex trees.

The prediction function at iteration t is updated as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad \text{……..(3)}$$

Where f_t is the function learned at the t -th step (e.g., a decision tree), and $\hat{y}_i$ is the predicted value for sample i . The objective function to be minimized at each step consists of a differentiable loss function l , measuring the difference between the predicted and true values, and a regularization term $\Omega(f)$, that penalizes model complexity:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t) \quad \text{……..(4)}$$

The regularization term helps prevent overfitting and is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad \text{…….(5)}$$

where:

- T is the number of leaves in the decision tree,
- w_j is the weight assigned to leaf j
- γ and λ are regularization hyperparameters.

By using a second-order Taylor approximation of the loss function, XGBoost achieves efficient optimization and supports parallel computation, making it well-suited for large-scale and high-dimensional datasets.

In the context of crop type classification using global satellite data (such as WorldCereal), XGBoost demonstrated excellent performance by capturing non-linear interactions between features like crop type, irrigation status, seasonality, and vegetation indices. Its robustness to noise and missing values further strengthens its applicability in precision agriculture, particularly for heterogeneous, real-world geospatial datasets [5][6][7].

Principal Component Analysis (PCA) is employed in this study as a dimensionality reduction technique and for the purpose of visualizing high-dimensional data in a two-dimensional space. The goal of PCA is to project the original dataset X into a lower-dimensional space Z , while preserving as much variance as possible. Mathematically, the transformation can be expressed as:

$$Z = XW \quad \text{……(6)}$$

Where:

$X \in R^{n \times d}$ is the original data matrix,

$W \in R^{d \times k}$ is the matrix of eigenvectors (principal components) of the covariance matrix of X

$Z \in R^{n \times k}$ - is the transformed representation in the reduced k - dimensional space

PCA enables the identification of latent patterns and variance structures in the data, which is particularly valuable for exploring class separability, assessing data clustering tendencies, and facilitating visualization.

K-means Clustering is an unsupervised learning algorithm that partitions the dataset into K clusters by minimizing the within-cluster sum of squared distances between data points and their corresponding cluster centroids. The optimization objective is defined as:

$$\min_{\mu_1,\ldots,\mu_K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \qquad \text{……..(7)}$$

Where:

- μ_k is the centroid of cluster C_k ,
- C_k is a set of instance given in that cluster .

In this study, K -means is applied in the PCA-reduced feature space, allowing the discovery of agro-climatic zones with similar geospatial and spectral characteristics, independent of any prior labeling. This unsupervised clustering contributes to the identification of regional patterns and structural groupings in agricultural land cover.

The original dataset used in this study contained geographical coordinates (lon, lat), crop labels (Crop), administrative boundaries (Country, State), and regional classifications (Region, Subregion, Continent, aez_id). All features were used either for exploratory data analysis, visualization, or filtering. Additional derived variables were generated during preprocessing, including numerical indices and NDVI-based estimates.

During exploratory data analysis, a set of visualizations was generated to assess feature distributions across crop classes. In Figure 1, we show the boxplot of a selected numeric variable, derived during feature extraction, across all labeled crop categories. This visualization helps identify inconsistencies, outliers, and heterogeneity within individual classes. Such intra-class variability is known to impair classifier accuracy, especially in multi-class settings, where clear feature separability is essential for reliable prediction.

RESULTS AND DISCUSSION

In the analysis conducted on the *world_cereals_scaled_final_all.csv* dataset — a preprocessed and feature-engineered subset derived from the WorldCereal: Global Crop Type Product (ESA/VITO, 2021) — a series of data preparation and modeling steps were carried out to classify crop types using global satellite indices and categorical attributes. (https://www.streambatch.io/knowledge/esa-worldcereal-global-crop-monitoring-at-field-scale). The original dataset was filtered to retain only high-probability cropland areas with valid crop type labels and enriched with externally computed NDVI features. Dimensionality reduction (PCA), statistical visualization (boxplots and violin plots), and supervised learning models (Random Forest, XGBoost) were then applied. Model performance was evaluated using confusion matrices and standard classification metrics.

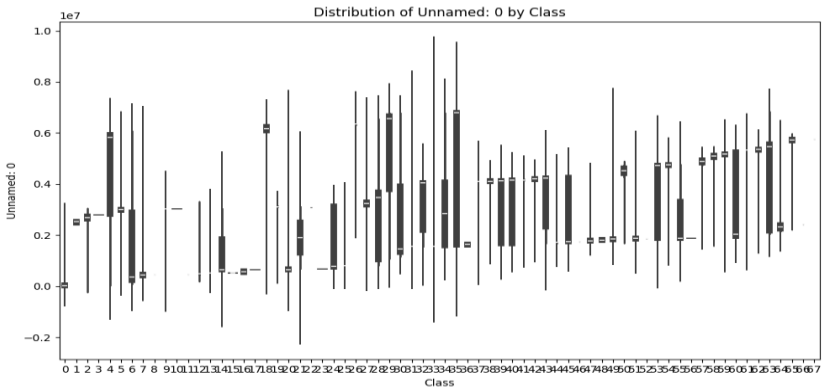

Figure 1. Boxplot distribution of a selected numerical feature across crop classes. The variable on the y-axis (Unnamed: 0) represents a numerical attribute derived from the WorldCereal dataset, associated with either NDVI-derived vegetation indices or a spatial indicator (depending on preprocessing pipeline version). Each boxplot summarizes the distribution of that feature within one crop class (x-axis), showing median, interquartile range, and outliers. High intra-class variance and the presence of outliers in several crop categories suggest heterogeneity in satellite signal representation for those crops, which may reflect mixed land cover, mislabeled samples, or seasonal overlap. These factors can negatively affect classifier performance by introducing ambiguity and reducing class separability.

The boxplot illustrates the distribution of feature values (likely locational or NDVI-related) across different classes, which represent either crop types or geographical entities. Considerable intra-class variability and the presence of outliers suggest heterogeneity within certain crop categories or possible mislabeling of samples. Such variance can negatively impact classifier performance and indicates the need for further class filtering or merging.

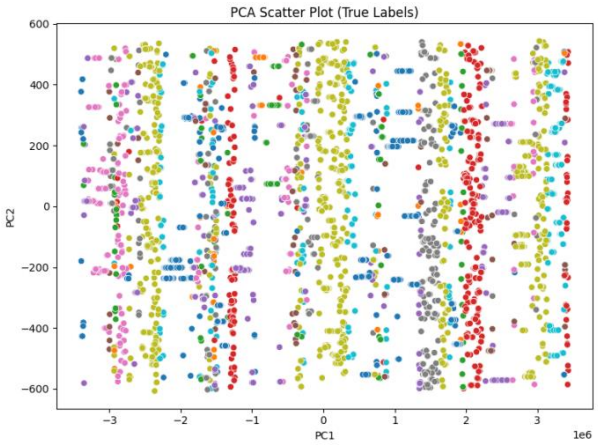

Figure 2. PCA scatter plot of crop type instances, colored by true class labels. Data was reduced to two principal components using PCA on standardized features from the WorldCereal dataset. The figure illustrates class structure and potential overlaps.

The PCA-based scatter plot (Figure 2) displays data instances projected onto the first two principal components, derived from standardized numerical features obtained from the WorldCereal dataset. The analysis was performed using Scikit-learn's PCA implementation in Python, and the resulting plot was generated by the authors using Matplotlib. While some class overlap is present — which is expected due to the high dimensionality and redundancy in spectral features — distinct groupings and isolated clusters can also be observed. These indicate that certain crop types exhibit unique spectral or temporal patterns, and that dimensionality reduction can reveal latent structures in the data, facilitating preliminary class separation and supporting downstream classification tasks.

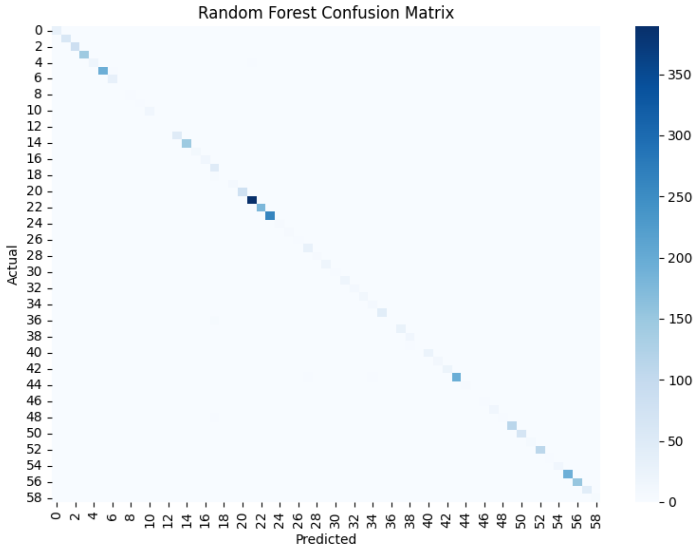

Figure 3. Confusion matrix for Random Forest classifier. Generated from the classification results on the test subset of the WorldCereal dataset using a Random Forest model implemented in Scikit-learn. Actual classes are shown on the y-axis, predicted classeson the x-axis. Color intensity represents the number of instances per class pair.

Figure 3 presents the confusion matrix for the Random Forest classifier trained on the preprocessed WorldCereal dataset. The matrix summarizes model performance across more than 60 crop classes and was generated using Scikit-learn, based on predictions on a held-out test set comprising 30% of the data.

The matrix demonstrates strong alignment along the diagonal, indicating high classification accuracy for the majority of classes. Misclassifications occur primarily between spectrally similar crop types, as also observed in the PCA projection.

Importantly, class distribution was sufficiently balanced to minimize model bias. The Random Forest model proved robust in handling high-dimensional and partially redundant datasets, leveraging its ensemble nature to effectively learn complex class boundaries. Higher misclassification rates were observed in rare classes with limited samples and overlapping spectral signatures.

The Random Forest classifier achieved an overall accuracy of 0.99 and a weighted F1-score of 0.98, demonstrating high precision and well-balanced performance across most classes. Major crop classes with large sample sizes (e.g., Classes 28, 29, 30, 50, 63, 64) achieved near-perfect classification results, with F1-scores ≥ 0.99. This underscores the model's ability to learn complex patterns and discriminate dominant crop types, even under high feature dimensionality.

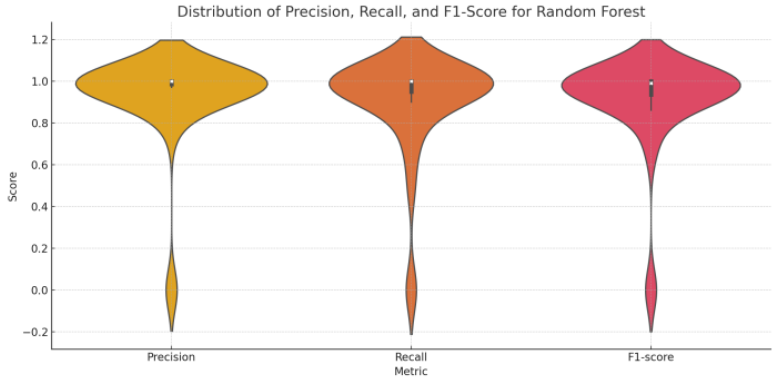

Figure 4. Violin plot of Random Forest performance metrics. To further investigate class-wise performance, we computed precision, recall, and F1-score for each crop class using predictions on the test set. Figure 4 shows the distribution of these metrics via violin plots. While the majority of classes achieved high precision and recall, a small number of minority classes showed significantly lower scores. This distributional view emphasizes the challenges in classifying underrepresented classes, despite the overall high average accuracy.

Despite these high scores, reduced performance was noted for minority and rare classes—especially Classes 25, 43, and 46—which had F1-scores of 0.00, indicating complete misclassification. These challenges stem from limited sample sizes and potential feature overlap with dominant classes. However, their impact on the overall weighted score is minimal due to their low representation.

The XGBoost model (Figure 5) delivered comparable overall performance to Random Forest, achieving a test set accuracy of 0.99 and a weighted F1-score of 0.99. However, macro-averaged precision (0.91) and recall (0.87) were slightly lower, reflecting greater variability across classes.

Figure 5 shows violin plots of the per-class precision, recall, and F1-score values obtained from predictions on the test set. These metrics were computed using classification_report from Scikit-learn, and the plot was generated in Seaborn. The distribution shapes indicate that while XGBoost performed well for dominant crop classes, a subset of minority classes exhibited low recall or F1-score — in some cases, being entirely missed (e.g., Classes 13, 15, 17).

XGBoost operates as a gradient boosting algorithm that sequentially adds decision trees to minimize a regularized loss function. Each tree in the ensemble focuses on correcting the residuals of the previous ones. Although it offers strong regularization and efficiency, the model tends to optimize toward well-represented classes unless explicit class weighting or sampling is applied.

This behavior is evident in the compressed lower tails of the violin plots, which represent underperforming or misclassified crop types.

One advantage of XGBoost lies in its stronger regularization and reduced risk of overfitting. However, it was observed to "ignore" underrepresented classes more frequently, especially when such classes lacked strong representation in tree splits.

While the overall accuracy appears impressive, per-class evaluation highlights the issue of sample imbalance, which can lead to the marginalization of low-frequency classes. This is a typical problem in multiclass datasets with skewed class distributions and should be addressed through additional techniques such as SMOTE (Synthetic Minority Oversampling Technique), weighted loss functions, or aggregation of similar classes.

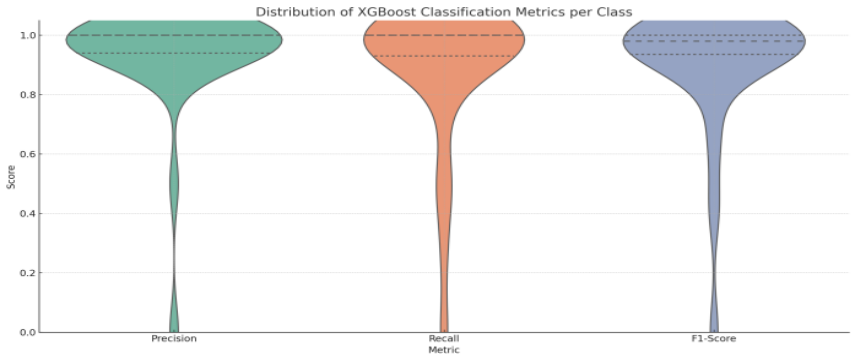

Figure 5. XGBoost performance – violin plot. The distribution reflects performance variability across crop types. Metrics were computed using Scikit-learn based on test set predictions, and the plot was generated using Seaborn.

Table 1. Overall performance comparison of Random Forest and XGBoost classifiers based on test set predictions. Metrics computed using Scikit-learn.

Model	Accuracy	Macro F_1	Weighted F_1
Random Forest	0.99	0.91	0.98
XGBoost	0.99	0.88	0.99

Table 1 summarizes the classification results of both models on the test set. Metrics were computed using Scikit-learn's classification_report function. Accuracy refers to the overall proportion of correct predictions. Macro F1 represents the unweighted mean F1-score across all classes, while Weighted F1 accounts for class imbalance by weighting each class's F1-score by its support. These results confirm that both models performed well, with slight differences in class-wise stability.

Both models demonstrated high overall accuracy, yet they differed in performance patterns across classes. Random Forest exhibited greater stability when dealing with rare or underrepresented classes, owing to its inherent ability to distribute attention evenly across samples. In contrast, XGBoost achieved superior performance in well-represented classes, attaining an exceptionally high weighted F1-score, but at the cost of diminished recall for minority classes.

This discrepancy underscores the importance of selecting classification algorithms in accordance with the class distribution and intended use case. While Random Forest ensures more balanced classification, XGBoost can provide sharper delineation for dominant patterns, especially in large-scale datasets.
These findings highlight the trade-offs involved in ensemble model selection and emphasize the need for careful model evaluation depending on dataset imbalance and application objectives [1].

CONCLUSION

This study experimentally demonstrated the applicability of ensemble-based machine learning models—namely Random Forest and XGBoost—for the classification of crop types and the identification of agro-ecological zones using globally available satellite data.

By utilizing the open-access WorldCereal dataset (ESA, 2021), and applying tailored preprocessing, it was possible to classify over 60 crop categories with exceptionally high accuracy (F1-score ≥ 0.98 for major classes).

The models were trained and evaluated on a representative subset of global cropland areas, with input features including crop type, seasonality, irrigation status, and derived NDVI indices. Despite the complexity of the multi-class setting and the presence of underrepresented categories, Random Forest showed greater robustness to class imbalance, while XGBoost delivered sharper delineation for dominant crop types.
Additionally, the use of Principal Component Analysis (PCA) for feature space reduction and K-means clustering for unsupervised agro-zone identification confirmed the potential to group agriculturally similar regions without prior labeling. This opens perspectives for developing self-organizing systems for land use monitoring, especially valuable in regions lacking extensive ground-truth data.

Importantly, the methodology proposed in this research is fully implementable using Python and relies exclusively on publicly accessible data, making it suitable for low-resource settings. The framework is not limited to crop type classification alone, and can be extended to related domains such as irrigation optimization, stress detection, soil monitoring, or yield prediction.

Overall, the study provides a reproducible, scalable, and data-driven approach for supporting precision agriculture through machine learning, with direct implications for regions with limited local infrastructure and growing needs for sustainable agricultural practices.

REFERENCES

[1] Butsko Christina, Van Tricht, K., Tseng, G., Milli, G., Rolnick, D., Cartuyvels, R., Becker-Reshef, I., Szantoi, Z. & Kerner, H. Deploying. 2025. Geospatial Foundation Models in the Real World: Lessons from WorldCereal. In *TerraBytes-ICML 2025 Workshop*. DOI:10.48550/arXiv.2508.00858.

[2] Franch, B., Cintas, J., Becker-Reshef, I., Sanchez-Torres, M. J., Roger, J., Skakun, S.,Sobrino, J.A., Tricht, K.V., Degerickx, J., Gilliams, S., Koetz, B., Szantoi, Z., & Whitcraft, A. 2022. Global crop calendars of maize and wheat in the framework of the WorldCereal project. *GIScience & Remote Sensing*, *59*(1), pp.885-913.

[3] Lemenkova, P. 2021. Distance-based vegetation indices computed by SAGA GIS: A comparison of the perpendicular and transformed soil adjusted approaches for the LANDSAT TM image. *Poljoprivredna tehnika*, *46*(3). pp.49-60.
[4] Luo, K., Lu, L., Xie, Y., Chen, F., Yin, F., & Li, Q. 2023. Crop type mapping in the central part of the North China Plain using Sentinel-2 time series and machine learning. *Computers and Electronics in Agriculture*, *205*, Article 107577.
[5] Maponya, M. G., Van Niekerk, A., Mashimbye, Z. E. 2020. Pre-harvest classification of crop types using a Sentinel-2 time-series and machine learning. *Computers and Electronics in Agriculture*, *169*, Article 105164.
[6] Mulla, D. J. 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4), pp.358-371.
[7] Sabir, R. M., Mehmood, K., Sarwar, A., Safdar, M., Muhammad, N. E., Gul, N., Athar, F., Danish Majeed, M., Sattar, J., Khan, Z., Fatima, M., Hussain, M.A., & Akram, H. M. B. 2024. Remote sensing and precision agriculture: a sustainable future. In *Transforming agricultural management for a sustainable future: climate change and machine learning perspectives.* Cham: Springer Nature Switzerland. pp. 75-103.
[8] Segarra, J. 2024. Satellite imagery in precision agriculture. In *Digital agriculture: a solution for sustainable food and nutritional security* Cham: Springer Internat. Publishing. pp. 325-340.
[9] Tubiello, F.N., Fabi, C., Conchedda, G., Casse, L. & Bottini, G. 2025. Comparison of WorldCereal with FAOSTAT data: an exploratory analysis. FAO Statistics Working Paper Series, No. 25-46. Rome, FAO. https://doi.org/10.4060/cd4154en.
[10] Van Tricht, K., Degerickx, J., Gilliams, S., Zanaga, D., Battude, M., Grosu, A., Beombacher, J., Lesiv, M., Laso Bayas, J.C., Karaman, S., Fritz, S., Becker-Reshef, I., Franch, B., Molla-Bononad, B., Boogaard, H., Kumar Pratihast, A., Koetz, B., & Szantoi, Z. 2023. WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, *15*(12), pp.5491-5515.
[11] Wang, Y., Zhang, Z., Feng, L., Ma, Y., & Du, Q. 2021. A new attention-based CNN approach for crop mapping using time series Sentinel-2 images. *Computers and Electronics in Agriculture*, *184*, Article 106090.

PRIMENA GLOBALNIH SATELITSKIH PODATAKA I MODELA MAŠINSKOG UČENJA ZA KLASIFIKACIJU USEVA I IDENTIFIKACIJU AGRO-ZONA

Nataša S. Milosavljević[1], **Rade Radojević**[1]

[1]*University of Belgrade, Faculty of Agriculture, Institute of Agricultural Engineering, Belgrade-Zemun, Republic of Serbia*

Abstrakt: Ova studija predstavlja primenjen pristup korišćenju globalno dostupnih satelitskih podataka, konkretno WorldCereal baze – za klasifikaciju i prostornu analizu poljoprivrednog zemljišta. Dataset, razvijen u okviru Evropske svemirske agencije (ESA), sadrži informacije o tipu useva, verovatnoći obradivosti zemljišta, sezonalnosti i statusu navodnjavanja za 2021. godinu, u rezoluciji od 10 metara. Nakon filtriranja na osnovu visoke verovatnoće poljoprivredne upotrebe, implementirani su algoritmi Random Forest i XGBoost za klasifikaciju tipova useva, kao i nenadgledano K-means klasterovanje za identifikaciju agro-zona sličnih karakteristika. Glavne kategorije useva uključuju žitarice, kukuruz, pirinač i mahunarke. Metodologija je testirana na velikom podskupu globalnog skupa podataka, koristeći Python-baziranu analitičku obradu.

Naši rezultati pokazuju da modeli mašinskog učenja mogu postići visoku tačnost klasifikacije ($F1 \geq 0.98$) čak i uz ograničen broj ulaznih parametara, dok klaster analiza otkriva konzistentne prostorne obrasce bez prethodnog označavanja.
Ovi nalazi ističu potencijal otvorenih satelitskih podataka i modelovanja zasnovanog na podacima u razvoju ekonomičnih i skalabilnih sistema za preciznu poljoprivredu, naročito u regionima bez detaljnih podataka sa terena.

Ključne reči: *daljinska detekcija, klasifikacija pokrivenosti zemljišta, stabla algoritama učenja, analiza prostornih podataka, mapiranje navodnjavanja, mašinsko učenje*.

Prijavljen:
***Submitted*: 29.07.2025.**
Ispravljen:
***Revised*: 01.08.2025.**
Prihvaćen:
***Accepted*:25.08.2025.**