eISSN 2783-6851

VILNIUS TECH
Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2025
Volume 3
Issue 1
Pages 1–17
https://doi.org/10.3846/ntcs.2025.23701

ENHANCING LAND USE CLASSIFICATION WITH HYBRID MACHINE LEARNING AND SATELLITE IMAGERY

Julius JANCEVIČIUS

Department of Information Systems, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania

Article History:
- received 10 April 2025
- accepted 14 May 2025

Abstract. The growing accessibility of satellite imagery and the rapid evolution of machine learning (ML) techniques have significantly advanced land use classification for environmental monitoring. However, challenges such as cloud coverage, varying image resolutions, and seasonal changes continue to hinder classification accuracy and consistency. This study aims to improve land use classification by proposing an integrated cloud interpolation, vegetation indices and ML based approach for classification of Sentinel-2 (S2) satellite data across the Baltic States. Specifically, a spatiotemporal interpolation module is introduced that reconstructs cloud-obscured pixels using multi-temporal coherence and derives optimized vegetation-index composites to enhance class separability under varying seasonal conditions. In order to achieve this aim and to choose the best ML algorithm for land use classification, we compare the performance of three classification algorithms, i.e., Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM), and evaluate their effectiveness in handling noisy and incomplete data. Our experimental results show that all three methods achieve strong classification accuracy, with RF exceeding 90%, while KNN and SVM also demonstrate competitive results. These methodological enhancements have been demonstrated to reduce cloud-induced misclassification and provide a scalable, transferable framework for operational land-use mapping in challenging atmospheric and seasonal contexts. These findings highlight the robustness of the proposed approach and provide valuable insights for future applications of ML in land use classification and environmental analysis.

Keywords: Sentinel-2, land use classification, image recognition, Random Forest, Support Vector Machine, machine learning, cloud interpolation.

Corresponding author. E-mail: *julius.jancevicius@stud.vilniustech.lt*

1. Introduction

The rapid advancement of remote sensing technologies, coupled with the increasing availability of high-resolution satellite imagery, has greatly enhanced the potential for efficient monitoring and management of land resources. However, land use classification remains a significant challenge due to the variability of spectral signatures, seasonal dynamics, and frequent cloud cover, which often obscure satellite imagery (Anandakrishnan et al., 2024; Rodríguez-Puerta et al., 2024). Conventional classification techniques often struggle to maintain high accuracy under such dynamic and uncertain conditions, thereby necessitating the adoption of more sophisticated approaches based on artificial intelligence (Marchetti et al., 2022). Among the various machine learning techniques applied to land use classification, RF, SVM, and KNN have emerged as widely used and empirically validated algorithms due to their respective strengths in handling complex, high-dimensional, and often noisy remote sensing data.

RF, an ensemble-based method, has been widely used in a variety of geographic contexts and has been praised for its robustness, interpretability, and ability to integrate multi-source data such as Sentinel-1 SAR and Sentinel-2 optical imagery (Dobrinić et al., 2021). Its effectiveness in dealing with heterogeneity and partial cloud cover has been demonstrated in numerous studies through strategies such as cloud masking, multi-temporal compositing, and spectral index fusion (Y. Wang et al., 2024; Xue et al., 2023). SVM, a margin-based classifier, has gained recognition for its strong performance in high-dimensional feature spaces, particularly when class distributions are non-linear or unbalanced. Research has shown that SVM can deliver competitive results, especially when used in conjunction with object-oriented classification and cloud segmentation methods. Its ability to construct optimal decision boundaries with limited training samples makes it particularly useful in remote sensing applications where labelled data may be scarce or unevenly distributed (Flohr et al., 2021; Hejmanowska & Kramarczyk, 2025). While conceptually simpler, KNN remains a valuable baseline in remote sensing due to its non-parametric nature and adaptability to different data distributions. Although often more sensitive to noise and computationally demanding on large datasets, KNN has shown solid performance when combined with effective preprocessing steps, including spectral filtering and cloud interpolation (Bebie et al., 2022; Pokhariya et al., 2023; Souza & Rodrigues, 2023).

Since each of these algorithms has distinct advantages and limitations when dealing with cloudy imagery and varying land cover types, a comparative analysis is essential to determine their relative effectiveness under real-world remote sensing conditions. This study, therefore, focuses on evaluating and comparing RF, SVM, and KNN for land use classification using Sentinel-2 imagery over Lithuania, a region characterised by diverse land use categories, seasonal variability, and persistent cloud presence. By systematically investigating their classification accuracy and robustness in the presence of cloud-related data challenges, the current study aims to identify the most reliable approach for practical applications in environmental monitoring.

Consequently, this paper presents a hybrid approach that integrates machine learning, cloud interpolation, and vegetation indices to improve land cover classification. Focusing on Lithuania as a case study, we have developed a robust classification pipeline that includes satellite data acquisition, data pre-processing (e.g., cloud interpolation and spectral index calculation), and classification using ML algorithms.

The main contributions and novelties of this work are:

1. A refined pre-processing workflow that improves classification under cloudy conditions.
2. Improved feature selection using vegetation indices for better separation of land use classes.
3. An evaluation of the accuracy results obtained by different classification algorithms to determine their respective strengths and limitations.

The rest of the paper is structured as follows. Section 2 (Related works) provides a detailed review of recent studies that using ML algorithms for land cover classification, highlighting the effectiveness of the RF, SVM, and KNN algorithms in different contexts. Section 3 (Methodology) outlines the proposed hybrid approach to land use classification using Sentinel-2 imagery. This section is divided into data acquisition, pre-processing, classifica-

tion, and post-processing stages, with each step in the pipeline described in detail. Section 4 (Experimental results) presents the results of applying the proposed approach to Sentinel-2 data over Lithuania. Section 5 (Discussion) examines the results, discussing the effectiveness and limitations of the approach. Section 6 (Conclusions and future works) summarises the main findings and contributions of the study.

2. Related works

The utilization of ML techniques in remote sensing has increasingly gained traction, with scholars concentrating on diverse facets of land cover and landscape pattern analysis to tackle challenges related to ecology and urban growth. Recent scholarly work underscores the adaptability and strength of these techniques in managing intricate spatial data across varied environments. The following discussion presents an analysis of pertinent studies on land cover classification.

For instance, in snow cover detection, Sentinel-2 imagery excels over Landsat, particularly in monitoring the dynamics of snow in mountainous areas like Switzerland, thanks to its superior spatial and temporal resolutions (Poussin et al., 2025). Moreover, in the northern Congo Republic, Sentinel-2 demonstrated outstanding performance in land use and cover mapping, achieving an overall accuracy of 93.80% and a Kappa coefficient of 0.89, which surpasses the 91.60% accuracy and 0.85 Kappa achieved by Landsat 9 (Bill Donatien et al., 2024). Additionally, the finer spatial resolution of Sentinel-2 mitigates mixed pixel issues and shows greater consistency with in-situ spectral measurements during field comparisons (Trevisiol et al., 2024). This finding indicates that Sentinel-2 satellite data is more appropriate for use in the classification task.

In addressing this issue, it is imperative to ascertain the provenance of the data. In many cases, erroneous choices in the initial step can have a direct negative impact on the results. Drawing upon the authors' extensive experience in this domain, it is evident that the utilisation of Sentinel-2 satellite data consistently yields more accurate results. Consequently, the subsequent literature analysis will be grounded in sources that evaluate the outcomes of classification algorithms when utilizing Sentinel-2 satellite imagery (Asmiwyati et al., 2025; Chanev et al., 2025; Chi et al., 2025; Rynkiewicz et al., 2025).

Cloud interpolation in remote sensing is crucial for filling data gaps in satellite imagery due to clouds. Recent methods like the spatial-spectral random forest (SSRF) use adjacent and similar pixels to effectively remove cloud interference by integrating spatial and spectral data. Additionally, deep learning techniques like partial convolution in a U-Net architecture have notably enhanced land surface temperature data interpolation, achieving a 44% decrease in root mean square error over traditional methods. These innovations are vital for improving the accuracy of satellite-derived environmental data (Patel et al., 2024; Q. Wang et al., 2022; Zhou et al., 2022).

For a more comprehensive overview of the used methods and outcomes from various studies employing ML algorithms in remote sensing, some relevant related works are compared below in Table 1. A general review of the literature was conducted to identify which machine-learning algorithms and accuracy metrics are most commonly applied to Sentinel-2

Table 1. Comparison of the related works found on land use classification

Reference	Algorithm	Accuracy metrics	Most Accurate Algorithm
(1)	(2)	(3)	(4)
Souza and Rodrigues (2023)	KNN*, DT*, SVM*, RF*	KIC*, OA*	KNN
Pokhariya et al. (2023)	DT, KNN, ANN*, SVM, BDT*, RF	KIC, OA	RF
Aliabad et al. (2022)	KNN, SVM, DT, RF	KIC, OA	KNN
Bebie et al. (2022)	RF, KNN, BR*	RMSE*, R^{2*}	RF
Kycko et al. (2022)	SVM, RF	OA, KIC, F1*	RF
Ren et al. (2023)	SVM, RF, KNN, LR*	RMSE, R^2, MAE*	RF
Kluczek et al. (2023)	RF, SVM	F1	RF
Yu et al. (2023)	RF, KNN, SVM, ANN	KIC, F1	ANN
Kamenova et al. (2024)	SVM, RF	F1	SVM
Kluczek et al. (2024)	RF, SVM, XGB*	OA	RF
Zhang et al. (2023)	KNN, SVM, RF	KIC, OA	KNN

Note: * K-Nearest Neighbour (KNN), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Boosted Decision Trees (BDT), Boosting Regressions (BR), Artificial Neural Network (ANN), Lasso Regression (LR), XGBoost (XGB), Overall Accuracy (OA), Kappa Index/Coefficient (KIC), Determination Coefficient (R^2), Root Mean Square Error (RMSE), F1-Score (F1), Mean Absolute Error (MAE).

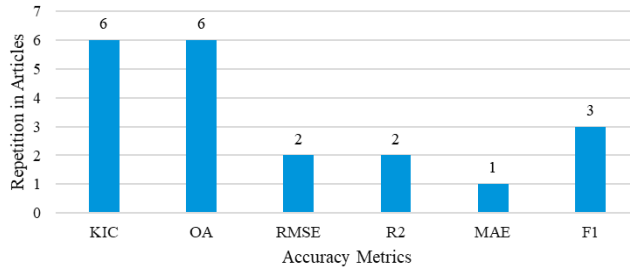


Figure 1. Accuracy metrics employed by the authors

imagery. The selection of articles was based on prevailing methodological trends and their direct relevance to the research objectives of the study, ensuring the inclusion of studies that offer significant empirical and conceptual contributions. Table 1 provides an organized comparison of the algorithms used, the accuracy metrics evaluated, and the most accurate algorithm of investigation.

Following a thorough analysis of the articles selected for the literature review, which were found to be highly correlated with the topic under consideration, it was observed that the classification algorithms (RF, KNN, and SVM) selected for the study (see column 2 of Table 1) were often grouped similarly in other scientific studies (Aliabad et al., 2022; Pokhariya et al., 2023; Ren et al., 2023; Souza & Rodrigues, 2023; Yu et al., 2023). Notwithstanding, the most accurate algorithm (in accordance with the results presented in the comparative articles) was identified as RF (see column 4 of Table 1), which was most efficient in six out of eleven

analysed articles (Bebie et al., 2022; Kluczek et al., 2023, 2024; Kycko et al., 2022; Pokhariya et al., 2023; Ren et al., 2023) Conversely, KNN in a superior position was successful in three out of eleven cases (Aliabad et al., 2022; Souza & Rodrigues, 2023; Zhang et al., 2023), while SVM laureate was only once (Kamenova et al., 2024). It has also been observed that the authors utilise a variety of reliability metrics. A review of the accuracy metrics employed by the authors is presented in Figure 1, which illustrates the popularity of the metrics.

As demonstrated in Figure 1, the examined sources demonstrated a predominant reliance on the OA and KIC accuracy metrics. On the other hand, the least popular accuracy metrics were RMSE, R2 and MAE, while the F1 accuracy metric is classified as moderately used.

3. Materials and methods

This section outlines the proposed hybrid ML, cloud interpolation, and vegetation index-based approach for classifying S2 satellite imagery (Figure 2). The process is divided into three main stages: data acquisition, pre-processing, and classification. In the data acquisition stage, S2 images are collected and filtered based on cloud cover and resolution parameters. The pre-processing stage prepares the data for analysis, including image merging, cloud removal, cloud interpolation and calculation of spectral indices to improve feature representation. At the classification stage, RF, KNN, and SVM models are used to assign land use classes using training datasets created from accurately labelled reference data. Each step in this approach is designed to improve the accuracy and reliability of the classification results (Stachura et al., 2024).

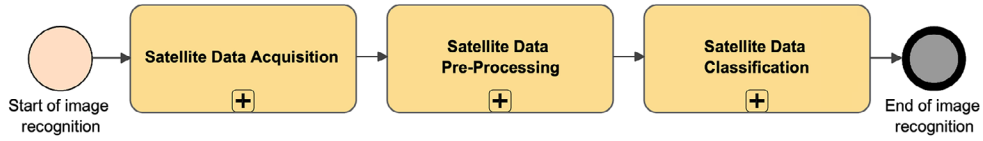


Figure 2. A schema of the hybrid ML, cloud interpolation, and vegetation indices-based approach

Each subprocess denoted by a plus symbol (+) in Figure 2 is elaborated upon in the following subsections.

3.1. Satellite data acquisition

The initial stage of our approach entails the acquisition of Sentinel-2 satellite imagery. Managed by the European Space Agency (ESA), Sentinel-2 provides multi-spectral data with resolutions ranging from 10 meters to 60 meters. To ensure optimal classification accuracy, Level-2A data, which is atmospherically corrected and pre-processed to minimize noise, was utilized. The study concentrated on the territory of Lithuania, with Sentinel-2 data spanning from 2022 to 2024. To enhance the quality of data for classification purposes, stringent filtering criteria were applied, selecting only images with cloud coverage below 5%.

The process of downloading satellite data is ensured by using the developed download script and receiving search criteria requested from the user. The request returns the metadata

of the S2 satellite images, which is then displayed to the user, conveying the territorial coverage of the selected images. If the selected images are deemed suitable, the logic transitions to the request processing stage.

The download of the data is performed in four threads, which means that four different satellite images are sent simultaneously. Upon completion of the download process, the images are then prepared for pre-processing, which involves archiving the received files and deleting unnecessary layers of satellite images. The bands utilised are enumerated in Table 2.

Table 2. Sentinel-2 bands

Sentinel-2 band	Resolution
Band 1 (B1) – Coastal aerosol	60 m
Band 2 (B2) – Blue	10 m
Band 3 (B3) – Green	10 m
Band 4 (B4) – Red	10 m
Band 5 (B5) – Vegetation Red Edge	20 m
Band 6 (B6) – Vegetation Red Edge	20 m
Band 7 (B7) – Vegetation Red Edge	20 m
Band 8 (B8) – NIR	10 m
Band 8A (B8A) – Vegetation Red Edge	20 m
Band 9 (B9) – Water vapour	60 m
Band 10 (B10) – SWIR - Cirrus	60 m
Band 11(B11) – SWIR	20 m
Band 12 (B12) – SWIR	20 m

As indicated from Table 2, this study uses only 10- and 20-meters resolution layers. This decision was made in light of the desired data accuracy and the analysis performed, which demonstrated that 60 m. resolution layers did not contribute to the study's classification accuracy.

3.2. Satellite data pre-processing

Pre-processing is performed to enhance the quality of the images as follows:

- **Image Integration**: Sentinel-2 bands were amalgamated into a single multi-band composite to facilitate comprehensive spectral analysis (Fan et al., 2022; Lemenkova, 2022; Ole Ørka et al., 2013; Schürz et al., 2023).
- **Background Cleaning:** NoData pixels and irrelevant background noise were removed using a thresholding approach (Juhász et al., 2023; Logan et al., 2024).
- **Compression:** To optimize memory resources, the DEFLATE compression method is applied to pre-processed images (Gonzalez et al., 2024; Jeromel & Žalik, 2020; Kai & Yuxiang, 2024).
- **Cloud Removal & Interpolation:** An advanced cloud detection algorithm utilizing the Scene Classification Layer (SCL) was employed to identify and mask cloud-covered areas. Subsequently, interpolation techniques were applied to reconstruct missing data, thereby ensuring the continuity and integrity of the classification process (Liu et al., 2021; Psychalas et al., 2023; Shao & Zou, 2021; Shepherd et al., 2020).

- **Spectral Index Calculation:** Three spectral indices – Normalized Difference Tillage Index (NDTI), Normalized Difference Vegetation Index Red-Edge (NDVIre), and Modified Normalized Difference Water Index (MNDWI) – were computed to enhance feature differentiation (Belayhun et al., 2024; Casamitjana et al., 2020; Farhadi et al., 2024; Ioannou, 2023; Lee et al., 2024; Niazmardi et al., 2018; Sankaran et al., 2023; Terzi Türk & Balçik, 2023).

In order to elucidate the modus operandi of cloud interpolation in greater detail, the process is initiated through the iteration of each S2 tile (territorial unit). It is important to note that from the available satellite images of the selected period (less than a month), it is known to which tile each satellite image belongs. Within the later iteration, for each territorial unit, the lowest cloud cover image is selected (cloud cover information is obtained from the satellite image metadata) and read as the parent image. A new iteration cycle is initiated, marking the onset of a second iteration. This cycle traverses the remaining tile images (child images), meticulously checking if the parent image contains NoData pixels that could potentially be filled by the child images.

Adding, in the context of working with Sentinel-2 imagery, it is not uncommon to encounter tiles from multiple Universal Transverse Mercator (UTM) zones, particularly in instances where the area of interest traverses' longitudinal boundaries. Each S2 tile is assigned a label using the Military Grid Reference System (MGRS), where the initial first tile digits (e.g., 34 and 35) correspond to specific UTM zones. These zones utilize distinct Coordinate Reference Systems (CRS), with UTM Zone 34N employing European Petroleum Survey Group (EPSG) EPSG:32634 and Zone 35N utilising EPSG:32635, to name but two examples. Consequently, spatial misalignment may occur when attempting to merge, compare, or analyse tiles across different zones. To ensure spatial consistency, it is imperative that all tiles are reprojected into a common CRS prior to any processing. This can be achieved in one of two ways: either by projecting all data into a single UTM zone or by converting all tiles to a global reference system such as World Geodetic System (WGS) WGS 84 (EPSG:4326). Reprojection is a crucial pre-processing step that ensures accurate geospatial analysis, classification, and visualization across tile boundaries (Roy et al., 2016). It is imperative to note that this step is incorporated within the overall process of image merging.

Upon completion of the aforementioned steps, the satellite images are deemed suitable for classification. This ensures that the input data is spatially consistent, radiometrically corrected, and aligned to a uniform coordinate reference system, thereby minimizing classification errors and enhancing the reliability of analytical results.

3.3. Satellite data classification

In this study, classification models such as RF, SVM, and K-Means are used and trained on labelled datasets obtained from Sentinel-2 imagery to classify land use categories.

The models will undergo configuration and extensive parameter testing to facilitate a comparative analysis of the results generated by different classifiers. The preparation of the training dataset utilises government-provided land use records, thereby enhancing the accuracy of land use class identification across various plots within the Republic of Lithuania, and potentially minimising classification errors across different temporal periods.

The accuracy of the classification is measured using reliability metrics such as: *Cohen's Kappa* (*CK*) (Dobrinić et al., 2021; Y. Wang et al., 2024; Z. Wang, 2023), *precision* and *recall* (Eisfelder et al., 2024; Farhadiani et al., 2024), *F1-score* (Albertini et al., 2024; Farhadiani et al., 2024; Flohr et al., 2021), and finally *Overall Accuracy* (*OA*) (Albertini et al., 2024; Farhadiani et al., 2024; Xue et al., 2023). Each metric is expressed as a percentage ranging from 0 to 100. It is imperative to acknowledge that the primary emphasis will be directed towards the outcomes derived from the OA, CK and F1 metrics.

4. Results

This section delineates the experimental results achieved through the application of the proposed method for Sentinel-2 data classification. To facilitate the experiment, a prototype was developed, ensuring the method's procedural steps were systematically executed. The findings are segmented and presented in accordance with each procedural step of the method.

4.1. Satellite data acquisition

During this stage, around 100 GB of S2 satellite data were methodically collected from ESA packages, which included satellite images in JPEG2000 (.jp2) file format. It is evident that the study utilised a total of 30 to 40 satellite images, which pixels were employed for the training of the machine learning (ML) algorithms. For the purpose of classification, 5 satellite images from different seasons (spring, summer or autumn) were selected, which have the largest coverage of the Republic of Lithuania and is characterised by land use change.

Each of the Sentinel-2 bands was individually downloaded and securely stored on a local storage system. After the successful completion of the download, all archives were extracted, and superfluous files were eliminated to refine the dataset.

The dataset comprises observations from the year 2023, effectively capturing seasonal variations across the Lithuanian territory. The analysed Sentinel-2 tiles were systematically categorized into the ethnographic regions of Lithuania, facilitating more precise territorial traceability. The details of the analysed tiles are presented in Table 3.

Table 3. Sentinel-2 tiles used in the experiment

Region	Sentinel-2 Tiles
Žemaitija	34UDG, 34VDH, 34UEG, 34VEH
Aukštaitija	34VFH, 34UFG, 35VLC, 35ULB, 35UMB, 35VMC
Suvalkija	34UFF, 34UFE, 35ULA, 34UGE
Dzūkija	35ULV, 35UMA, 35UMV

As demonstrated in Table 3, a combination of 17 tiles is required for comprehensive coverage of the Republic of Lithuania. It is important to note that Lithuania falls within two satellite orbits, which are otherwise called UTM zones (see section 2.2). Specifically, the satellite images (to cover Lithuania's territory) obtained originate from zones 34 and 35. For illustrative purposes, a 35VLD tile instance has been provided, which depicts the capital of Latvia, Riga (see Figure 3).

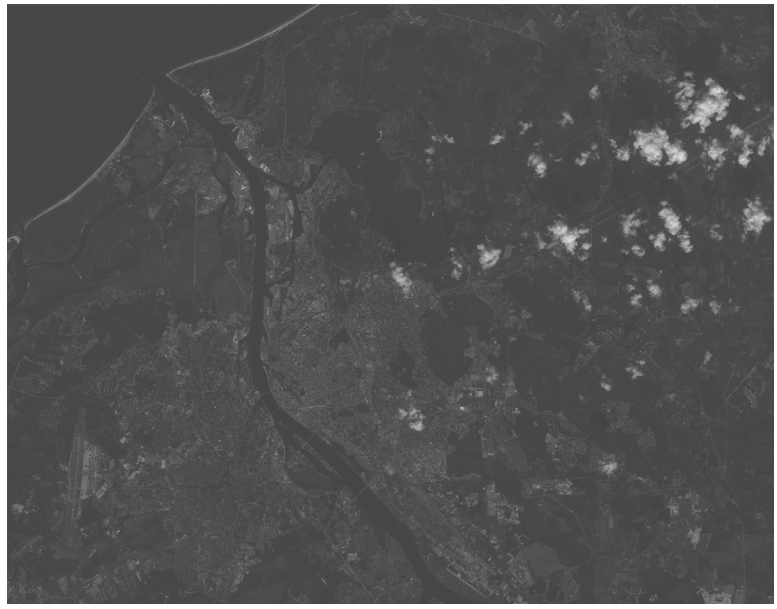


Figure 3. Cropped 35VLD tile B04 band (red) image (2024-09-05)

4.2. Satellite data pre-processing

The outcomes of the pre-processing stage are inherently difficult to quantify through direct testing or measurement. However, several key enhancements are implemented during this stage, including the integration of individual satellite bands into a composite image, the calculation of spectral indices, the removal of cloud-covered pixels using the Sentinel-2 Scene Classification Layer (SCL), and subsequent interpolation to fill cloud-related gaps. These procedures are designed to maximize image quality and ensure data continuity across temporal observations.

Conversely, it can be inferred that the applied compression techniques on the merged satellite data are effective. The original multi-band satellite images, comprising all 13 Sentinel-2 bands, consistently require approximately 3.13 GB of storage. However, following compression, the memory footprint is significantly reduced – by nearly half for a full, pixel-covered image – demonstrating the efficiency of the applied method. A detailed comparison of the compression results is presented in Figure 4.

1. 20240905 T35VMC 0.0%.tiff	3.13 GB
2. 20240905 T35ULA 0.05%.tiff	3.13 GB
3. 20240905 T35UMB 0.03%.tiff	3.13 GB
4. 20240905 T34VFH 0.07%.tiff	3.13 GB
6. 20240905 T35VLC 0.09%.tiff	3.13 GB
7. 20240905 T35UMA 0.0%.tiff	3.13 GB
8. 20240905 T35ULB 0.19%.tiff	3.13 GB

a)

1. 20240905 T35VMC 0.0%.tiff	1.5 GB
2. 20240905 T35ULA 0.05%.tiff	1.8 GB
3. 20240905 T35UMB 0.03%.tiff	889 MB
4. 20240905 T34VFH 0.07%.tiff	1.85 GB
6. 20240905 T35VLC 0.09%.tiff	1.84 GB
7. 20240905 T35UMA 0.0%.tiff	288.8 MB
8. 20240905 T35ULB 0.19%.tiff	1.84 GB

b)

Figure 4. The outcome of satellite image compression: a – original satellite image sizes prior to compression; b – satellite image sizes post-compression

Figure 5. Cropped 35ULA tile merged satellite image (2024-08-28)

An example of a pre-processed satellite image prepared for classification, corresponding to tile 35ULA – which encompasses the capital city of Lithuania, Vilnius – is presented in Figure 5.

Given these comprehensive measures, it can be concluded that the pre-processing outputs are of high quality and effectively align with the initial goals established at the onset of the project.

4.3. Satellite data classification

A notable outcome of the study was the observation that incorporating cloud interpolation markedly enhanced the consistency of classification results in regions frequently affected by cloud cover. By leveraging temporal information to reconstruct missing data, the model was able to sustain high levels of accuracy despite adverse atmospheric conditions. This approach enhancement effectively mitigated the impact of cloud obstruction, thereby increasing the robustness and operational applicability of the classification process for land monitoring purposes.

Note that within the context of this study, cloud interpolation served solely as a supplementary pre-processing technique to enhance data quality prior to classification. The primary objective is to evaluate classification algorithms, identify the most accurate model, and determine the optimal set of hyperparameters that maximize performance.

The hyperparameters themselves, for each algorithm, were identified based on a literature review, where the hyperparameter values were identified in a test study using GridSearchCV, where possible combinations of hyperparameters are presented and the best combination is selected (Ahmad et al., 2022; Alshammari, 2024; Vazirani et al., 2024).

The optimal hyperparameters for each classification algorithm are presented in Table 4.

Table 4. Optimal hyperparameters for each classification algorithm

Classifier	Optimal hyperparameters
RF	*n_estimators = 100, max_depth = 20, min_samples_leaf = 4, min_samples_split = 2*
SVM	*C = 0.1, gamma = scale, kernel = rbf*
KNN	*n_neighbors = 10, weights = uniform, p = 2*

The hyperparameter optimization process yielded distinct optimal configurations for each classification algorithm. For the KNN model, the most effective setup involved using 10 neighbours, uniform distance weighting, and the Euclidean distance metric, corresponding to $p = 2$. This configuration suggests that the classifier benefits from considering a moderately sized local neighbourhood while treating all neighbours equally, regardless of their proximity. In the case of the RF algorithm, the optimal parameters included the use of 100 decision trees, a maximum depth of 20, a minimum of four samples required at each leaf node, and a minimum of two samples required to split an internal node. Bootstrapping was enabled, and no additional class weighting was applied, indicating a preference for a moderately deep, unweighted ensemble structure. The SVM achieved its best configuration using a radial basis function (RBF) kernel, with a regularization parameter C set to 0.1 and gamma set to "scale", suggesting a relatively soft margin and automatic adjustment of the kernel coefficient based on the input data. These parameter selections reflect the specific modelling characteristics required to effectively capture the patterns within the dataset for each algorithm.

The classification accuracy achieved by each evaluated algorithm is summarized in Table 5.

Table 5. Classification performance metrics for evaluated algorithms

Classifier	Cohen's Kappa	F1-score	Recall	Precision
RF	89.23%	90.53%	90.86%	90.21%
SVM	86.23%	87.93%	88.16%	87.71%
KNN	84.73%	85.08%	85.36%	84.81%

The classification results (see Table 5) revealed distinct performance differences among the evaluated algorithms, reflecting the influence of both model architecture and optimized hyperparameter configurations. The RF classifier achieved the highest validation performance, demonstrating its robustness and ability to capture complex, non-linear relationships within the feature space through ensemble learning and deep tree structures. The SVM, employing an RBF kernel with a soft margin configuration, performed comparably well, indicating its effectiveness in handling high-dimensional data and separating classes with non-linear boundaries. While the KNN algorithm exhibited slightly lower performance, its simplicity and reliance on local neighbourhood structures still enabled it to deliver competitive results. Overall, the findings suggest that ensemble-based approaches, particularly RF, are well-suited for land cover classification tasks involving multi-spectral satellite imagery, while SVM and KNN remain valuable alternatives depending on the computational constraints and complexity of the data.

The land use classes identified are summarized in Figure 6. It should be noted that not all classes are identified during the relevant periods.

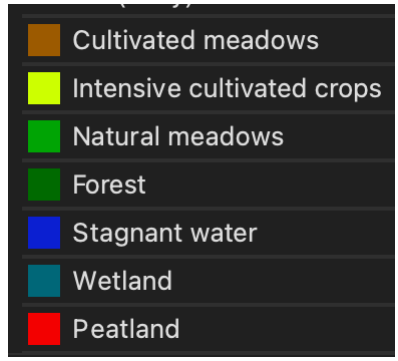
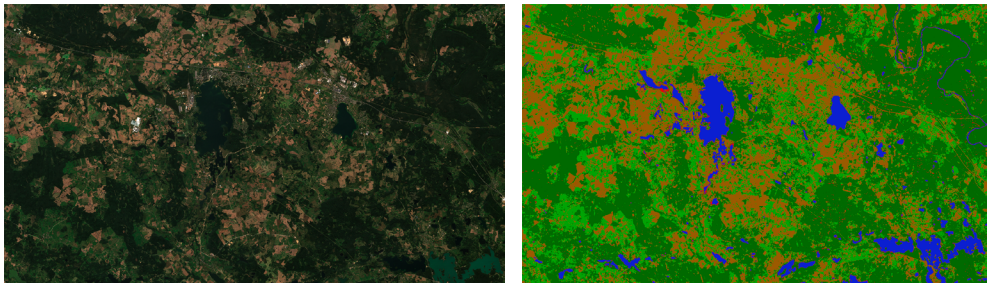


Figure 6. Identified land use classes

Examples of the classification result (using RF classifier) are shown in Figures 7 and Figure 8.

a) b)

Figure 7. Cropped tile (35ULA) example of classification result (2024-09-05): a – pre-processed input satellite image; b – output satellite image of classification algorithm



a) b)

Figure 8. Cropped tile (35ULA) example of classification result (2024-09-05): a – pre-processed input satellite image; b – output satellite image of classification algorithm

5. Discussion

This study primarily aimed to evaluate the efficacy of different machine learning algorithms in improving land use classification, using Sentinel-2 satellite imagery augmented by a sophisticated pre-processing framework that included cloud interpolation and the use of vegetation indices. The focus was to ascertain which algorithm among RF, KNN, and SVM performs best under the enhanced conditions provided by our hybrid approach.

Our analysis revealed distinct advantages and limitations of each algorithm, tailored by the pre-processing enhancements that were part of our hybrid approach. Random Forest emerged as the most effective, achieving superior classification accuracy. This outcome can be attributed to RF's inherent ability to handle large and complex datasets with high-dimensional features, making it particularly suited to the spectral diversity and the variabilities introduced by cloud cover and seasonal changes in satellite imagery.

Support Vector Machines demonstrated strong performance in scenarios with clear distinctions between classes due to its effective handling of high-dimensional space. However, its performance was slightly behind RF, suggesting that while SVM is robust, the RF's ensemble method provides a more adaptable and error-tolerant approach in the complex and varied environments typical of land use classification tasks.

K-Nearest Neighbours, while generally less robust in noisy environments like satellite data affected by cloud interference, benefited significantly from the preprocessing steps, particularly cloud interpolation which reduced noise in the input data. This preprocessing allowed KNN to perform competitively, highlighting its utility when conditions are optimized for its operation.

5.1. Limitations of the research

The study focused on a predefined set of well-known ML algorithms, potentially overlooking newer or unconventional methods that might offer improved results in land use classification. Additionally, while the pre-processing enhancements significantly improved algorithm performance, they also increased the computational load. This could be a constraint in operational settings where speed and efficiency are critical. Future research should explore more efficient pre-processing methods to balance accuracy with computational demands.

6. Conclusions and future works

The findings from this study support the hypothesis that preprocessing enhancements, coupled with robust machine learning algorithms, can significantly improve the accuracy of land use classification. Random Forest, in particular, stood out as the most effective algorithm under the conditions tested. These results not only reinforce the importance of choosing the right algorithm but also underscore the value of a comprehensive preprocessing regime that aligns with the specific strengths and weaknesses of the chosen classifiers.

Looking forward, it would be beneficial to expand this analysis by incorporating additional machine learning models, such as deep learning architectures, which might offer further improvements in classification accuracy. Additionally, exploring the integration of these algorithms in a real-time analysis framework could potentially enhance the applicability of this research in operational settings.

References

Ahmad, G. N., Fatima, H., Shaf, U., Salah Saidi, A., & Imdadullah. (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. *IEEE Access*, *10*, 80151–80173. https://doi.org/10.1109/ACCESS.2022.3165792

Albertini, C., Gioia, A., Iacobellis, V., Petropoulos, G. P., & Manfreda, S. (2024). Assessing multi-source random forest classification and robustness of predictor variables in flooded areas mapping. *Remote Sensing Applications: Society and Environment*, *35*, Article 101239. https://doi.org/10.1016/j.rsase.2024.101239

Aliabad, F. A., Malamiri, H. R. G., Shojaei, S., Sarsangi, A., Ferreira, C. S. S., & Kalantari, Z. (2022). Investigating the ability to identify new constructions in urban areas using images from unmanned aerial vehicles, Google Earth, and Sentinel-2. *Remote Sensing*, *14*(13), Article 3227. https://doi.org/10.3390/rs14133227

Alshammari, T. (2024). Using artificial neural networks with GridSearchCV for predicting indoor temperature in a smart home. *Engineering, Technology and Applied Science Research*, *14*(2), 13437–13443. https://doi.org/10.48084/etasr.7008

Anandakrishnan, J., Sundaram, V. M., & Paneer, P. (2024). CERMF-Net: A SAR-Optical feature fusion for cloud elimination from Sentinel-2 imagery using residual multiscale dilated network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 11741–11749. https://doi.org/10.1109/JSTARS.2024.3411032

Asmiwyati, I. G. A. A. R., Gargita, I. W. D., & Wiguna, P. P. K. (2025). Analysis of urban green open space development in North Denpasar District, Denpasar City, Bali, Indonesia. *Geographia Technica*, *20*(1), 79–96. https://doi.org/10.21163/GT_2025.201.07

Bebie, M., Cavalaris, C., & Kyparissis, A. (2022). Assessing Durum wheat yield through Sentinel-2 imagery: A machine learning approach. *Remote Sensing*, *14*(16), Article 3880. https://doi.org/10.3390/rs14163880

Belayhun, M., Chere, Z., Abay, N. G., Nicola, Y., & Asmamaw, A. (2024). Spatiotemporal pattern of water hyacinth (Pontederia crassipes) distribution in Lake Tana, Ethiopia, using a random forest machine learning model. *Frontiers in Environmental Science*, *12*, Article 1476014. https://doi.org/10.3389/fenvs.2024.1476014

Bill Donatien, L. M., Biona Clobite, B., & Lemvo Meris Midel, M. (2024). Comparing Sentinel-2 and Landsat 9 for land use and land cover mapping assessment in the north of Congo Republic: A case study in Sangha region. *International Journal of Remote Sensing*, *45*(2), 8015–8036. https://doi.org/10.1080/01431161.2024.2394238

Casamitjana, M., Torres-Madroñero, M. C., Bernal-Riobo, J., & Varga, D. (2020). Soil moisture analysis by means of multispectral images according to land use and spatial resolution on andosols in the colombian andes. *Applied Sciences*, *10*(16), Article 5540. https://doi.org/10.3390/app10165540

Chanev, M., Kamenova, I., Dimitrov, P., & Filchev, L. (2025). Evaluation of Sentinel-2 deep resolution 3.0 data for winter crop identification and organic barley yield prediction. *Remote Sensing*, *17*(6), Article 957. https://doi.org/10.3390/rs17060957

Chi, Z., Chen, H., Chang, S., Li, Z. L., Ma, L., Hu, T., Xu, K., & Zhao, Z. (2025). Large-Scale Monitoring of potatoes late blight using multi-source time-series data and Google Earth engine. *Remote Sensing*, *17*(6), Article 978. https://doi.org/10.3390/rs17060978

Dobrinić, D., Gašparović, M., & Medak, D. (2021). Sentinel-1 and 2 time-series for vegetation mapping using random forest classification: A case study of Northern Croatia. *Remote Sensing*, *13*(12), Article 2321. https://doi.org/10.3390/rs13122321

Eisfelder, C., Boemke, B., Gessner, U., Sogno, P., Alemu, G., Hailu, R., Mesmer, C., & Huth, J. (2024). Cropland and crop type classification with Sentinel-1 and Sentinel-2 time series using Google Earth Engine for agricultural monitoring in Ethiopia. *Remote Sensing*, *16*(5), Article 866. https://doi.org/10.3390/rs16050866

Fan, Z., Zhan, T., Gao, Z., Li, R., Liu, Y., Zhang, L., Jin, Z., & Xu, S. (2022). Land cover classification of resources survey remote sensing images based on segmentation model. *IEEE Access*, *10*, 56267–56281. https://doi.org/10.1109/ACCESS.2022.3175978

Farhadi, H., Ebadi, H., Kiani, A., & Asgary, A. (2024). Near real-time flood monitoring using multi-sensor optical imagery and machine learning by GEE: An automatic feature-based multi-class classification approach. *Remote Sensing*, *16*(23), Article 4454. https://doi.org/10.3390/rs16234454

Farhadiani, R., Homayouni, S., Bhattacharya, A., & Mahdianpari, M. (2024). Crop classification using multi-temporal RADARSAT constellation mission compact polarimetry SAR data. *Canadian Journal of Remote Sensing*, *50*(1), Article 2384883. https://doi.org/10.1080/07038992.2024.2384883

Flohr, P., Bradbury, J., & ten Harkel, L. (2021). Tracing the patterns: Fields, villages, and burial places in Lebanon. *Levant*, *53*(3), 315–335. https://doi.org/10.1080/00758914.2021.1968114

Gonzalez, S. T., Velez-Zea, A., & Barrera-Ramírez, J. F. (2024). High performance holographic video compression using spatio-temporal phase unwrapping. *Optics and Lasers in Engineering*, *181*, Article 108381. https://doi.org/10.1016/j.optlaseng.2024.108381

Hejmanowska, B., & Kramarczyk, P. (2025). Assessing land cover changes using the LUCAS database and sentinel imagery: A comparative analysis of accuracy metrics. *Applied Sciences*, *15*(1), Article 240. https://doi.org/10.3390/app15010240

Ioannou, K. (2023). On the identification of agroforestry application areas using object-oriented programming. *Agriculture*, *13*(1), Article 164. https://doi.org/10.3390/agriculture13010164

Jeromel, A., & Žalik, B. (2020). An efficient lossy cartoon image compression method. *Multimedia Tools and Applications*, *79*(1–2), 433–451. https://doi.org/10.1007/s11042-019-08126-7

Juhász, L., Xu, J., & Parkinson, R. W. (2023). Beyond the tide: A comprehensive guide to sea-level-rise inundation mapping using FOSS4G. *Geomatics*, *3*(4), 522–540. https://doi.org/10.3390/geomatics3040028

Kai, X., & Yuxiang, Z. (2024). Improving the performance of 3D image model compression based on optimized DEFLATE algorithm. *Scientific Reports*, *14*(1), Article 14899. https://doi.org/10.1038/s41598-024-65539-7

Kamenova, I., Chanev, M., Dimitrov, P., Filchev, L., Bonchev, B., Zhu, L., & Dong, Q. (2024). Crop type mapping and winter wheat yield prediction utilizing Sentinel-2: A case study from Upper Thracian Lowland, Bulgaria. *Remote Sensing*, *16*(7), Article 1144. https://doi.org/10.3390/rs16071144

Kluczek, M., Zagajewski, B., & Kycko, M. (2024). Combining multitemporal optical and radar satellite data for mapping the Tatra Mountains non-forest plant communities. *Remote Sensing*, *16*(8), Article 1451. https://doi.org/10.3390/rs16081451

Kluczek, M., Zagajewski, B., & Zwijacz-Kozica, T. (2023). Mountain tree species mapping using Sentinel-2, PlanetScope, and Airborne HySpex hyperspectral imagery. *Remote Sensing*, *15*(3), Article 844. https://doi.org/10.3390/rs15030844

Kycko, M., Zagajewski, B., Kluczek, M., Tardà, A., Pineda, L., Palà, V., & Corbera, J. (2022). Sentinel-2 and AISA airborne hyperspectral images for Mediterranean Shrubland Mapping in Catalonia. *Remote Sensing*, *14*(21), Article 5531. https://doi.org/10.3390/rs14215531

Lee, J., Kim, K., & Lee, K. (2024). Multi-Sensor image classification using the random forest algorithm in Google Earth engine with KOMPSAT-3/5 and CAS500-1 images. *Remote Sensing*, *16*(24), Article 4622. https://doi.org/10.3390/rs16244622

Lemenkova, P. (2022). GRASS GIS scripts for Satellite image analysis by raster calculations using modules r.mapcalc, d.rgb, r.slope.aspect. *Tehnicki Vjesnik*, *29*(6), 1956–1963. https://doi.org/10.17559/TV-20220322091846

Liu, C., Huang, H., Hui, F., Zhang, Z., & Cheng, X. (2021). Fine-resolution mapping of pan-arctic lake ice-off phenology based on dense Sentinel-2 time series data. *Remote Sensing*, *13*(14), Article 2742. https://doi.org/10.3390/rs13142742

Logan, T. L., Smyth, M. M., & Calef, F. J. (2024). Planetary orbital mapping and mosaicking (POMM) integrated open source software environment. *Astronomy and Computing*, *46*, Article 100788. https://doi.org/10.1016/j.ascom.2024.100788

Marchetti, G., Bizzi, S., Belletti, B., Lastoria, B., Comiti, F., & Carbonneau, P. E. (2022). Mapping riverbed sediment size from Sentinel-2 satellite data. *Earth Surface Processes and Landforms*, *47*(10), 2544–2559. https://doi.org/10.1002/esp.5394

Niazmardi, S., Homayouni, S., Safari, A., McNairn, H., Shang, J., & Beckett, K. (2018). Histogram-based spatio-temporal feature classification of vegetation indices time-series for crop mapping. *International*

Journal of Applied Earth Observation and Geoinformation, *72*, 34–41. https://doi.org/10.1016/j.jag.2018.05.014

Ole Ørka, H., Gailis, J., Vege, M., Gobakken, T., & Hauglund, K. (2013). Analysis-ready satellite data mosaics from Landsat and Sentinel-2 imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(5), 2088–2101. https://doi.org/10.1109/JSTARS.2012.2228167

Patel, P. N., Jiang, J. H., Gautam, R., Gadhavi, H., Kalashnikova, O., Garay, M. J., Gao, L., Xu, F., & Omar, A. (2024). A remote sensing algorithm for vertically resolved cloud condensation nuclei number concentrations from airborne and spaceborne lidar observations. *Atmospheric Chemistry and Physics*, *24*(5), 2861–2883. https://doi.org/10.5194/acp-24-2861-2024

Pokhariya, H. S., Singh, D. P., & Prakash, R. (2023). Evaluation of different machine learning algorithms for LULC classification in heterogeneous landscape by using remote sensing and GIS techniques. *Engineering Research Express*, *5*(4), Article 045052. https://doi.org/10.1088/2631-8695/acfa64

Poussin, C., Peduzzi, P., & Giuliani, G. (2025). Snow observation from space: An approach to improving snow cover detection using four decades of Landsat and Sentinel-2 imageries across Switzerland. *Science of Remote Sensing*, *11*, Article 100182. https://doi.org/10.1016/j.srs.2024.100182

Psychalas, C., Vlachos, K., Moumtzidou, A., Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2023). Towards a paradigm shift on mapping muddy waters with Sentinel-2 using machine learning. *Sustainability*, *15*(18), Article 13441. https://doi.org/10.3390/su151813441

Ren, C., Jiang, H., Xi, Y., Liu, P., & Li, H. (2023). Quantifying temperate forest diversity by integrating GEDI LiDAR and multi-temporal Sentinel-2 imagery. *Remote Sensing*, *15*(2), Article 375. https://doi.org/10.3390/rs15020375

Rodríguez-Puerta, F., Perroy, R. L., Barrera, C., Price, J. P., & García-Pascual, B. (2024). Five-year evaluation of Sentinel-2 cloud-free mosaic generation under varied cloud cover conditions in Hawai'i. *Remote Sensing*, *16*(24), Article 4791. https://doi.org/10.3390/rs16244791

Roy, D. P., Li, J., Zhang, H. K., & Yan, L. (2016). Best practices for the reprojection and resampling of Sentinel-2 multi spectral instrument level 1C data. *Remote Sensing Letters*, *7*(11), 1023–1032. https://doi.org/10.1080/2150704X.2016.1212419

Rynkiewicz, A., Hościło, A., Aune-Lundberg, L., Nilsen, A. B., & Lewandowska, A. (2025). Detection and quantification of vegetation losses with Sentinel-2 images using bi-temporal analysis of spectral indices and transferable random forest model. *Remote Sensing*, *17*(6), Article 979. https://doi.org/10.3390/rs17060979

Sankaran, R., Al-Khayat, J. A., J, A., Chatting, M. E., Sadooni, F. N., & Al-Kuwari, H. A. S. (2023). Retrieval of suspended sediment concentration (SSC) in the Arabian Gulf water of arid region by Sentinel-2 data. *Science of the Total Environment*, *904*, Article 166875. https://doi.org/10.1016/j.scitotenv.2023.166875

Schürz, M., Grigoropoulou, A., García Márquez, J., Torres-Cambas, Y., Tomiczek, T., Floury, M., Bremerich, V., Schürz, C., Amatulli, G., Grossart, H. P., & Domisch, S. (2023). hydrographr: An R package for scalable hydrographic data processing. *Methods in Ecology and Evolution*, *14*(12), 2953–2963. https://doi.org/10.1111/2041-210X.14226

Shao, M., & Zou, Y. (2021). Multi-spectral cloud detection based on a multi-dimensional and multi-grained dense cascade forest. *Journal of Applied Remote Sensing*, *15*(02), Article 028507. https://doi.org/10.1117/1.jrs.15.028507

Shepherd, J. D., Schindler, J., & Dymond, J. R. (2020). Automated mosaicking of Sentinel-2 satellite imagery. *Remote Sensing*, *12*(22), Article 3680.. https://doi.org/10.3390/rs12223680

Souza, F. E. S. de, & Rodrigues, J. I. de J. (2023). Evaluation of machine learning algorithms in the classification of multispectral images from the Sentinel-2A/2B orbital sensor for mapping the environmental dynamics of Ria Formosa (Algarve, Portugal). *ISPRS International Journal of Geo-Information*, *12*(9), Article 361. https://doi.org/10.3390/ijgi12090361

Stachura, G., Ustrnul, Z., Sekuła, P., Bochenek, B., Kolonko, M., & Szczęch-Gajewska, M. (2024). Machine learning based post-processing of model-derived near-surface air temperature – A multimodel approach. *Quarterly Journal of the Royal Meteorological Society*, *150*(759), 618–631. https://doi.org/10.1002/qj.4613

Terzi Türk, S., & Balçik, F. (2023). Rastgele orman algoritması ve Sentinel-2 MSI ile fındık ekili alanların belirlenmesi: Piraziz Örneği. *Geomatik*, *8*(2), 91–98. https://doi.org/10.29128/geomatik.1127925

Trevisiol, F., Mandanici, E., Pagliarani, A., & Bitelli, G. (2024). Evaluation of Landsat-9 interoperability with Sentinel-2 and Landsat-8 over Europe and local comparison with field surveys. *ISPRS Journal of Photogrammetry and Remote Sensing*, *210*, 55–68. https://doi.org/10.1016/j.isprsjprs.2024.02.021

Vazirani, H., Wu, X., Srivastava, A., Dhar, D., & Pathak, D. (2024). Highly efficient JR optimization technique for solving prediction problem of soil organic carbon on large scale. *Sensors*, *24*(22), Article 7317. https://doi.org/10.3390/s24227317

Wang, Q., Wang, L., Zhu, X., Ge, Y., Tong, X., & Atkinson, P. M. (2022). Remote sensing image gap filling based on spatial-spectral random forests. *Science of Remote Sensing*, *5,* Article 100048. https://doi.org/10.1016/j.srs.2022.100048

Wang, Y., Jin, S., & Dardanelli, G. (2024). Vegetation classification and evaluation of yancheng coastal wetlands based on random forest algorithm from Sentinel-2 Images. *Remote Sensing*, *16*(7), Article 1124. https://doi.org/10.3390/rs16071124

Wang, Z. (2023). Spatial differentiation characteristics of rural areas based on machine learning and GIS statistical analysis – A case study of Yongtai County, Fuzhou City. *Sustainability*, *15*(5), Article 4367. https://doi.org/10.3390/su15054367

Xue, H., Xu, X., Zhu, Q., Yang, G., Long, H., Li, H., Yang, X., Zhang, J., Yang, Y., Xu, S., Yang, M., & Li, Y. (2023). Object-oriented crop classification using time series sentinel images from Google Earth Engine. *Remote Sensing*, *15*(5), Article 1353. https://doi.org/10.3390/rs15051353

Yu, H., Luo, Z., Wang, L., Ding, X., & Wang, S. (2023). Improving the accuracy of flood susceptibility prediction by combining machine learning models and the expanded flood inventory data. *Remote Sensing*, *15*(14), Article 3601. https://doi.org/10.3390/rs15143601

Zhang, H., He, J., Chen, S., Zhan, Y., Bai, Y., & Qin, Y. (2023). Comparing three methods of selecting training samples in supervised classification of multispectral remote sensing images. *Sensors*, *23*(20), Article 8530. https://doi.org/10.3390/s23208530

Zhou, J., Luo, X., Rong, W., & Xu, H. (2022). Cloud removal for optical remote sensing imagery using distortion coding network combined with compound loss functions. *Remote Sensing*, *14*(14), Article 3452. https://doi.org/10.3390/rs14143452