Proceedings of the 21st Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 365–373
DOI: 10.15439/2026F9420
ISSN 2300-5963 ACSIS, Vol. 47

Automated Detection of Sosnowsky's Hogweed Using Sentinel-2 Imagery and Machine Learning

Reinis Odītis, Ivo Odītis
Kārlis Freivalds
0009-0003-5488-1608, 0000-0003-2354-3780
0000-0003-2684-559X
University of Latvia
Raina Blvd. 19, Riga, Latvia
Email: ro17020@students.lu.lv, ivo.odits@lu.lv
karlis.freivalds@lu.lv

***Abstract*—Sosnowsky's hogweed (Heracleum sosnowskyi) is an invasive plant species widespread in Latvia, posing threats to biodiversity, human health, and economic activity. Manual re-inspection of the approximately 19,000 registered occurrence records is impractical at scale. This study proposes an automated machine learning framework for hogweed detection using freely available Sentinel-2 multispectral satellite imagery, with reference data sourced from the State Plant Protection Service geospatial database. Three supervised classification algorithms — Support Vector Machine, Random Forest, and Extreme Gradient Boosting — were evaluated across multiple dataset configurations combining spectral bands, derived indices, and multi-temporal features. Binary classification achieved a peak accuracy of 92.6%, while multi-class classification reached 87.7% (Cohen's kappa 0.836). The results demonstrate that Sentinel-2 imagery combined with multi-temporal features enables reliable automated monitoring of hogweed distribution at national scale.**

I. Introduction

INVASIVE plant species pose a serious threat to biodiversity by outcompeting native species, altering ecosystem structure, and degrading habitat quality for both wildlife and agricultural land. The uncontrolled spread of invasive species causes significant economic losses — both direct (crop damage, infrastructure harm) and indirect (management and eradication costs) — making timely and accurate distribution monitoring a prerequisite for effective control policy.

In this study, the authors seek to find an effective automated method for identifying and monitoring the distribution of Sosnowsky's hogweed (Heracleum sosnowskyi, [1]), an invasive plant species widespread in Latvia. Several approaches are examined, of which the use of Sentinel-2 satellite data is identified as the most suitable for accomplishing this particular task. Publicly available data from the State Plant Protection Service database on territories where hogweed is present, together with satellite-derived data and their spectral indices, are used to prepare datasets for model training and testing.

The authors evaluate three most suited to the task machine learning classification algorithms — Support Vector Machine, Random Forest, and Extreme Gradient Boosting — and compare their results. It is concluded that SVM achieved the highest overall accuracy among all evaluated algorithms. The outcome is a method that enables the identification hogweed stands in a given territory according to satellite data.

The structure of this paper is as follows:

- Section 2 explains the problem being addressed: the need for systematic annual monitoring of a large number of territories;
- Section 3 focuses on different remote sensing technologies and describes why satellite data are most applicable for this research;
- Section 4 presents the authors' proposed method, including reference data collection, data acquisition and preprocessing, selected classification algorithms, and related aspects;
- Section 5 discusses the results obtained and presents the conclusions.

II. Problem statement

A. Monitoring of Hogweed in Latvia

Sosnowsky's hogweed (Figure 1) was introduced into the territory of Latvia during the Soviet era as a promising fodder crop [2]. It is a species non-native to Latvian nature that threatens local plants and their habitats, causes economic losses, harm to the environment, and damage to human health. The State Plant Protection Service (SPPS) carries out the determination of the spread of Sosnowsky's hogweed and the surveying of invaded territories, providing data to municipalities for the coordination and planning of control measures within their administrative territories, and to the public, so that everyone is informed about the locations where the plant is spreading and can protect themselves from the burns it causes [3].

SRRS maintains the invasive plant species distribution database of the state information system for crop monitoring: territorial units and the invasive plant species present within them are recorded, along with their registration date and other attributes. Registration of new territories occurs manually, either following applications submitted by landowners or other residents. The reported territory is inspected and, if the presence of an invasive plant species is confirmed, the relevant



Fig. 1. Heracleum sosnovskyi

information is entered into the database. Problems arise with the repeated inspection of territories already registered in the database: this is a time-consuming task (approximately 19000 records for hogweed are currently registered) and is difficult to perform manually.

The aim of the study is to find a method that would allow the automated re-inspection of registered territories, preferably using publicly available data (for example, aerial photographs or satellite imagery). If the method proves sufficiently effective, it may even become possible to identify new territories of invasive species automatically.

B. Hogweed as a Research Subject

The research object identified by the authors of this study is hogweed. In the Latvian context, this encompasses several hogweed species; however, there are no significant morphological or phenological differences among them that would affect their monitoring — that is, although multiple hogweed species exist, they are morphologically and phenologically very similar. As Jana Müllerová and co-authors note in their publication [4], all hogweed species exhibit a structurally distinctive form, comprising stems two to five metres in height, leaves that can reach up to two metres in width, and clearly distinguishable white compound umbel inflorescences. The authors further emphasise that hogweed is most readily identifiable during the flowering phase, whereas outside this phase the spectral contrast between hogweed and other plant species decreases substantially.

To conduct research on invasive plant species using remote sensing technologies, it is necessary to analyse whether a given species possesses specific characteristics that could distinguish it from other plant species in the region. Hogweed species are well suited to this task by virtue of their morphological and phenological properties, as well as their spectral distinctiveness during the flowering period, which differs markedly from that of other plant species.

III. Remote sensing

A. Remote Sensing Options for Plant Monitoring

Several remote sensing technology alternatives exist, offering different types of acquirable data, varying levels of precision, and other distinct characteristics. The authors of [5] divide all remote sensing solutions into three groups: satellites, manned aviation, and unmanned aerial vehicles (UAVs).

The use of UAVs for plant species identification and classification has gained considerable popularity. This approach offers high spatial resolution and easily adaptable temporal resolution. UAVs are capable of conducting observations below cloud cover, which eliminates a particularly problematic factor, and can be customised by mounting specific sensors tailored to the requirements of a given problem. The primary limitation of UAVs is their constraint to surveying limited areas. When a study requires the coverage of large surfaces, data acquisition via UAVs is comparatively slow and dependent on battery capacity. A further significant drawback is that flights cannot be conducted in rain, snow, or other adverse weather conditions. It should also be noted that the use of UAVs may be subject to specific regulations depending on the country [6].

An older yet still relevant method is manned aviation, which enables aerial photography across various spectral ranges, providing very high spatial resolution and variable temporal resolution. Moreover, it allows for better coverage of large areas compared to UAVs [7].

Also worth mentioning from a plant identification perspective is a specific manned aviation service — *LiDAR*, or airborne laser scanning. This approach yields highly specific data that includes three-dimensional structural information. *LiDAR* can acquire data both during the day and at night, an capability that many other alternatives do not offer. This option carries limitations similar to those of aerial photography: the method is comparatively expensive, and although temporal resolution is theoretically adjustable, regular data acquisition entails disproportionately high costs in practice. It should additionally be noted that data acquired via *LiDAR* may be well suited to specific problems, but is insufficient on its own for plant species monitoring, as it does not include spectral information [8].

The principal drawback of manned aviation is its very high cost: both aircraft and pilot services are expensive. This approach is frequently used for one-time surveys, but may prove to be a poor solution for regular monitoring. In this context, however, it is worth noting that publicly available archives of aerial photographs exist, which may be useful for training machine learning models if such an approach is adopted [7].

Satellite remote sensing, in turn, provides global and regular coverage of the Earth's surface, making it an indispensable tool for large-scale vegetation monitoring. Satellite remote sensing is widely used for data acquisition at global and national scales, while for land cover or vegetation monitoring at the landscape scale, satellite data are generally sufficient for the

majority of studies [9]. At the same time, satellite systems have significant limitations compared to other platforms — they suffer from lower spatial resolution, cloud cover obstruction, and constraints related to positioning and revisit timing [10]. Nevertheless, despite these limitations, freely accessible satellite programmes such as the Copernicus Sentinel-2 offer considerable advantages. The principal strengths of Sentinel-2 are its high spatial resolution, high temporal resolution, and the availability of red-edge spectral bands, which together with open data access make it one of the most widely used sensors in vegetation research [11]. The following section includes a comparison of various satellite platforms.

B. Satellite Platforms

As noted above, remote sensing today offers a broad spectrum of data sources. Even when the selected plant species is suitable for identification using remote sensing technologies, it is important to choose a specific satellite platform appropriate for identifying that species. This subsection provides a comparison of the most widely used available satellite platforms, evaluated according to the following factors:

- **Spatial resolution**, indicating the smallest distinguishable object;
- **Temporal resolution**, indicating the cycle length after which the sensor will again acquire data at the same location;
- **Spectral bands**, indicating the range of electromagnetic radiation wavelengths observed;
- **Availability**, indicating whether the satellite platform is freely accessible, the extent of its coverage, and the size of the platform's archive, i.e., the date from which data are available.

The following satellite platforms are included in the comparison: *Landsat 8 [12], Landsat 9 [13], Sentinel-2 [14], MODIS [15], PlanetScope [16]*. The spatial resolution column in the table specifies the particular spectral bands to which each resolution value applies; a detailed description of spectral bands and their indices can be found in Subsection IV-C.

The satellite platform comparison is presented in Table I.

Examining the comparison presented in the table, it can be observed that the best spatial and temporal resolution is offered by the *PlanetScope* platform; however, it provides only four spectral bands and is commercially licensed, which poses serious difficulties for large-scale studies requiring regular data acquisition. The *MODIS* satellite platform has critically low spatial resolution, which is insufficient for detailed vegetation monitoring. Comparing the *Landsat 8/9* and *Sentinel-2* missions, similar overall performance can be observed; however, *Sentinel-2* offers superior spatial resolution, and its temporal resolution provides greater flexibility in the selection of data acquisition dates. This is particularly advantageous in situations of high cloud cover. In addition, *Sentinel-2* provides *red-edge* spectral bands, which may prove significant for the identification of specific plant species.

Based on the satellite platform comparison, the authors have selected the *Sentinel-2* satellite platform as the primary and sole data source for the development of the practical solution. It should also be noted here that Sentinel-2 acquires data more frequently than Landsat (Table I).

A critical aspect for non-commercial projects and research is data accessibility. When addressing problems such as vegetation or invasive plant monitoring, data are required on a regular basis, and their territorial coverage is important for consistent observations. The Sentinel-2 mission data are well suited to such situations, as they are freely accessible to anyone who requires them. The data are available through several platforms, including Google Earth Engine (GEE) [17], which the authors have selected for the implementation of the practical solution. GEE offers not only the ability to visually browse the data, but also an application programming interface (API) for automated data retrieval. As mentioned in the previous section, Sentinel-2 satellites acquire data in tiles of 100 × 100 km. The territory of Latvia is covered by four such tiles, meaning that complete coverage of Latvian territory is achievable [14].

One of the most significant limitations of Sentinel-2 data is spatial resolution, which for the most detailed spectral bands is 10 m. This means that the data granularity is insufficient for identifying small objects or, in the context of this work, individual hogweed stands. However, the spatial resolution is sufficient to identify the distribution of a plant species where its coverage exceeds the area of a single pixel (approximately 100 m^2).

IV. Proposed Method

This section presents a machine learning framework designed to identify *Heracleum sosnowskyi* using multi-temporal Sentinel-2 imagery in Latvia. We evaluate three supervised algorithms—Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—across four different dataset configurations, which vary in their spatial and temporal aspects. The performance of these algorithms is assessed using both binary and multi-class classification methods.

A. Study Area

This study was conducted using data from Latvia's (64,589 km^2) territory, where *H. sosnowskyi* is widespread, with over 19,000 documented occurrences [18]. The territory spans four Sentinel-2 tiles (34UEB, 34VEJ, 35UMA, 35VMC) in the UTM/WGS84 projection system [19], encompassing diverse land cover types, including forests (55%) and agricultural land (32%) [20].

B. Reference Data Collection

Reference data for *H. sosnowskyi* were obtained from the State Plant Protection Service (SPPS) geospatial database [18], which contains field-verified occurrence records across Latvia. The database stores polygon geometries with associated attributes including observation date (*Datums*), total infestation area (*Platība*), and dense coverage area (*Blīvā Platība*),

TABLE I
COMPARISON OF WIDELY USED SATELLITE PLATFORMS

Parameter	Landsat 8	Landsat 9	Sentinel-2	MODIS	PlanetScope
Spatial Resolution					
Visible/NIR	30 m	30 m	**10 m**	250–500 m	3 m
Red-edge	–	–	**20 m**	–	–
SWIR	30 m	30 m	20 m	500 m	–
Temporal Resolution					
Revisit cycle	16 days	16 days	**5 days**	Daily	Daily
Number of satellites	1	1	2	2	>200
Spectral Bands					
Total count	11	11	**13**	36	4
Red-edge bands	0	0	**3**	0	0
SWIR bands	2	2	2	5	0
Availability					
Data cost	Free	Free	**Free**	Free	Commercial
Coverage	Global	Global	**Global**	Global	Global
operational since	2013	2021	2015	1999	2016

representing the portion of each polygon with homogeneous hogweed dominance.

A three-stage filtering process was applied to ensure data quality and compatibility with Sentinel-2 spatial resolution. First, temporal filtering retained only records from the most recent five years, as SPPS performs manual field verification and newer records are more likely to represent current, actively maintained hogweed infestations. Second, occurrences with total area below 100 m² (0.01 ha) were excluded, as these fall below Sentinel-2's 10 m spatial resolution and cannot be reliably detected. Finally, to ensure high-quality training samples, 150 occurrences were selected based on the smallest difference between total polygon area and dense coverage area, indicating locations where hogweed forms homogeneous, dense stands. This threshold was determined empirically, as larger sample sizes significantly degraded data quality due to increased presence of mixed pixels and sparse vegetation patches.

To enable multi-class classification, negative reference samples representing non-hogweed land cover types were manually selected through visual interpretation using *geojson.io*. Six classes were defined: grassland, cropland, forest, urban, water, and bare soil, resulting in 300 total negative samples (Table II). Polygon centroids were computed for all hogweed occurrences to standardize the spatial representation, yielding 450 point coordinates for subsequent Sentinel-2 feature extraction.

TABLE II
REFERENCE SAMPLE DISTRIBUTION ACROSS LAND COVER CLASSES

Class	Samples	Percentage (%)
Hogweed	150	33.3
Grassland	135	30.0
Cropland	54	12.0
Forest	51	11.3
Urban	30	6.7
Water	18	4.0
Bare soil	12	2.7
Total	**450**	**100.0**

C. Data Acquisition and Preprocessing

Sentinel-2 Level-2A atmospherically corrected surface reflectance imagery was acquired via the AWS Element84 STAC API for all 450 reference point locations. Satellite imagery was acquired from the same year as each reference observation; negative class samples were manually collected from summer 2025 imagery. For each point, data acquisition targeted two phenological windows during the growing season: early summer (June 1-30) capturing pre-flowering vegetative growth, and peak summer (July 15 - August 31) corresponding to flowering and maximum biomass periods when hogweed exhibits distinct spectral characteristics [4]. A 100 m × 100 m bounding box was constructed around each point location, and available scenes with cloud cover below 30% were queried. The scene with the lowest cloud cover within each temporal window was selected for feature extraction.

Cloud and cloud shadow contamination were addressed using Sentinel-2's Scene Classification Layer (SCL), generated by the Sen2Cor atmospheric correction processor [14]. Pixels classified as cloud shadow (SCL=3), cloud medium probability (SCL=8), cloud high probability (SCL=9), or thin cirrus (SCL=10) were masked and excluded from the analysis. Additional quality filtering excluded pixels flagged as no data (SCL=0), saturated/defective (SCL=1), or snow/ice (SCL=11).

Bottom-of-Atmosphere (BOA) reflectance values were extracted for 10 spectral bands (Table III). Bands with 60 m native resolution (B1, B9, B10) were excluded, as B9 and B10 serve atmospheric correction purposes rather than surface characterization [14]. Multi-resolution alignment was achieved by resampling 20 m bands (B5-B7, B8A, B11-B12, SCL) to the 10 m grid via bilinear interpolation.

D. Feature Engineering and Dataset Configurations

Beyond the 10 Sentinel-2 spectral bands, seven spectral indices were calculated to introduce non-linear relationships (Table IV). These indices express specific spectral information such as vegetation health (NDVI, NDRE, EVI), water content (NDWI, MNDWI), bare soil (BSI), and built-up areas (NDBI).

TABLE III
SENTINEL-2 SPECTRAL BANDS

Band	Wavelength (nm)	Resolution (m)	Spectral Region
B2	490	10	Blue
B3	560	10	Green
B4	665	10	Red
B5	705	20	Red Edge 1
B6	740	20	Red Edge 2
B7	783	20	Red Edge 3
B8	842	10	Near-Infrared (NIR)
B8A	865	20	NIR Narrow
B11	1610	20	Shortwave Infrared 1
B12	2190	20	Shortwave Infrared 2

Four dataset configurations were constructed to evaluate how different information extraction approaches affect classification performance.

The single-pixel center (SP-center) configuration (17 features, 448 samples) extracted values from the center pixel of each 100 m × 100 m reference area, selecting the scene with lowest cloud coverage from either phenological window. This represents a traditional point-based sampling approach.

The single-pixel mean (SP-mean) configuration (17 features, 448 samples) utilized the same temporal selection strategy but extracted feature values as the spatial mean over all pixels within the reference area (approximately 100 pixels at 10 m resolution), reducing noise from individual pixel variability.

The multi-temporal (MT) configuration (48 features, 407 samples) incorporated features from both phenological windows using spatial averaging. All 17 base features were extracted separately for early summer and peak summer scenes, and phenological dynamics were captured through absolute and relative change calculations for seven key features (NDVI, NDRE, EVI, B08, B05, B06, MNDWI). The reduced sample size reflects the requirement for clear imagery in both temporal windows (cloud coverage $<30\%$).

The spatial 3×3 neighborhood (S3×3) configuration (68 features, 448 samples) integrated local spatial context by extracting a 3×3 pixel window from the center of each reference area and computing four statistics (mean, standard deviation, minimum, maximum) for each of the 17 base features.

TABLE IV
CALCULATED SPECTRAL INDICES

Index	Formula
NDVI	$\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}}$
NDRE	$\frac{\text{NIR}-\text{RedEdge}_1}{\text{NIR}+\text{RedEdge}_1}$
NDWI	$\frac{\text{NIR}-\text{SWIR}}{\text{NIR}+\text{SWIR}}$
MNDWI	$\frac{\text{Green}-\text{SWIR}}{\text{Green}+\text{SWIR}}$
BSI	$\frac{(\text{SWIR}+\text{Red})-(\text{NIR}+\text{Blue})}{(\text{SWIR}+\text{Red})+(\text{NIR}+\text{Blue})}$
NDBI	$\frac{\text{SWIR}-\text{NIR}}{\text{SWIR}+\text{NIR}}$
EVI	$2.5 \times \frac{\text{NIR}-\text{Red}}{\text{NIR}+6\times\text{Red}-7.5\times\text{Blue}+1}$

NIR: B8 (842 nm), Red: B4 (665 nm),
Green: B3 (560 nm), Blue: B2 (490 nm),
RedEdge_1: B5 (705 nm), SWIR: B11 (1610 nm)

E. Classification Algorithms and Experimental Design

Three machine learning algorithms were evaluated: Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Each classifier was trained in both binary classification (hogweed vs. non-hogweed) and multi-class classification (hogweed vs. six individual land cover types: grassland, cropland, forest, urban, water, bare soil).

Two hyperparameter optimization strategies were compared for each algorithm.

The default configuration used library-recommended parameters: SVM with $C = 1.0$, $\gamma = \text{scale}$, and RBF kernel; RF with 100 trees and unlimited depth; XGBoost with 100 boosting rounds and learning rate $\eta = 0.3$.

The grid search configuration employed 5-fold cross-validation to optimize key parameters: SVM explored $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$ with RBF kernel; RF tuned the number of trees (100-500), maximum depth (10-30 or unlimited), splitting criteria, and feature sampling strategies; XGBoost optimized boosting rounds(100-300), learning rate (0.01-0.3), tree depth (3-9), and regularization parameters.

The complete experimental matrix comprised 48 unique combinations: 4 dataset configurations (SP-center, SP-mean, MT, S3×3) × 3 algorithms (SVM, RF, XGBoost) × 2 classification tasks (binary, multi-class) × 2 hyperparameter strategies (default, grid search).

Data were partitioned into training (70%), validation (10%), and test (20%) sets using stratified sampling with a fixed random seed (42) across all experiments to ensure identical splits for fair comparison. The validation set was used for hyperparameter selection in grid search experiments, while final performance was evaluated on the held-out test set.

F. Implementation

Satellite data acquisition utilized the AWS Element84 STAC API for Sentinel-2 Level-2A imagery access, with rasterio 1.5.0 employed for Cloud-Optimized GeoTIFF windowed reading and bilinear resampling of 20 m bands to 10 m resolution. Vegetation indices were calculated using NumPy 2.4.4, and feature extraction was performed with pandas 3.0.2. Classification models were implemented in Python 3.13 using scikit-learn 1.8.0 (SVM, RF) and XGBoost 3.2.0. Model training and hyperparameter optimization leveraged scikit-learn's GridSearchCV with 5-fold cross-validation and parallel execution.

V. RESULTS

A. Overall Classification Performance

Tables V and VI summarize the best-performing models for each dataset configuration across binary and multi-class classification tasks. For each configuration, models were selected based on validation performance, with final results reported on the test set.

Binary classification achieved accuracies ranging from 88.9% (SP-center) to 92.6% (MT), with all configurations

exceeding 88%. The multi-temporal (MT) dataset combined with Support Vector Machine yielded the highest accuracy (92.6%) and a high ROC-AUC (0.952). The spatial 3×3 neighborhood (S3×3) configuration with XGBoost achieved a slightly higher ROC-AUC (0.958).

Multi-class classification, which distinguishes hogweed from six land cover classes (grassland, cropland, forest, urban, water, and bare soil), resulted in lower accuracies, ranging from 76.7% (SP-center) to 87.7% (MT). The multi-temporal configuration with XGBoost achieved the highest overall performance, with an accuracy of 87.7% and Cohen's kappa of 0.836.

All best-performing models employed hyperparameter optimization via grid search rather than default parameter settings.

TABLE V
BINARY CLASSIFICATION PERFORMANCE BY DATASET CONFIGURATION

Dataset	Algorithm	Tuning	Acc.	ROC-AUC
SP-center	RF	Grid	0.889	0.908
SP-mean	SVM	Grid	0.911	0.925
MT	SVM	Grid	0.926	0.952
S3×3	XGBoost	Grid	0.922	0.958

TABLE VI
MULTI-CLASS CLASSIFICATION PERFORMANCE BY DATASET CONFIGURATION

Dataset	Algorithm	Tuning	Acc.	Kappa
SP-center	SVM	Grid	0.767	0.695
SP-mean	SVM	Grid	0.833	0.784
MT	XGBoost	Grid	0.877	0.836
S3×3	SVM	Grid	0.833	0.784

B. Dataset Configuration Analysis

Figure 2 illustrates classification accuracy across all three algorithms and for both binary and multi-class tasks. Random Forest and XGBoost achieved their highest binary classification accuracy with the S3×3 configuration (91.1% and 92.2%, respectively), while the Support Vector Machine performed best with MT (92.6%). SP-mean outperformed SP-center in the evaluated configuration by 2.2% in binary classification (91.1% vs 88.9%).

Multi-class classification exhibited substantially lower accuracy than binary classification across all configurations, with performance gaps ranging from 4.9% for MT data to 12.2% for SP-center extraction. This performance reduction is likely due to the increased complexity of distinguishing hogweed from multiple land cover classes rather than from aggregated non-hogweed samples. The smallest performance gap for MT (92.6% binary vs 87.7% multi-class) suggests that phenological dynamics captured across two seasonal acquisition windows may provide discriminative information relevant to broader vegetation type classification beyond binary hogweed detection.

Figure 3 presents ROC curves for the best binary classification model per dataset configuration, providing deeper insight into discriminative performance beyond simple accuracy. All four configurations achieved ROC-AUC exceeding 0.90, with S3×3 achieving the highest ROC-AUC (0.958), followed closely by MT (AUC = 0.952).

C. Algorithm Performance Comparison

Tables VII, VIII, and IX present the best-performing configurations for each algorithm across all dataset configurations and both classification tasks.

Support Vector Machine (Table VII) achieved highest accuracy with the MT dataset (92.6%) using binary classification. Close performance was observed with SP-mean and S3×3 (both 91.1%) also using binary classification. Notably, SVM's best models primarily utilized grid search hyperparameter optimization, isuggesting that performance improvements were more frequently obtained through hyperparameter optimization compared to Random Forest and XGBoost.

TABLE VII
SUPPORT VECTOR MACHINE PERFORMANCE ACROSS DATASET CONFIGURATIONS

Dataset	Task	Method	Accuracy
SP-center	Binary	Default	0.867
SP-center	Multiclass	Grid	0.767
SP-mean	Binary	Grid	0.911
SP-mean	Multiclass	Grid	0.833
MT	Binary	Grid	0.926
MT	Multiclass	Grid	0.790
S3×3	Binary	Grid	0.911
S3×3	Multiclass	Grid	0.833

Random Forest (Table VIII) demonstrated robust performance with default parameters across most configurations, achieving binary classification accuracies ranging from 84.0% (MT) to 91.1% (S3×3). Multi-class accuracy remained stable at 74.4%–82.7%, with MT achieving the highest multi-class performance (82.7%) using default parameters.

TABLE VIII
RANDOM FOREST PERFORMANCE ACROSS DATASET CONFIGURATIONS

Dataset	Task	Method	Accuracy
SP-center	Binary	Grid	0.889
SP-center	Multiclass	Default	0.744
SP-mean	Binary	Default	0.889
SP-mean	Multiclass	Default	0.800
MT	Binary	Default	0.840
MT	Multiclass	Default	0.827
S3×3	Binary	Grid	0.911
S3×3	Multiclass	Default	0.800

XGBoost (Table IX) achieved the highest overall multi-class accuracy among all evaluated algorithms (87.7% with MT features using grid search). Binary classification peaked at 92.2% with S3×3 features. Grid search was required for optimal binary performance across all configurations, while its impact on multi-class performance varied depending on the dataset.

Overall, Random Forest exhibited the most stable performance with minimal tuning, Support Vector Machine benefited most consistently from hyperparameter optimization, and XGBoost achieved the highest peak performance, particularly for multi-class classification.

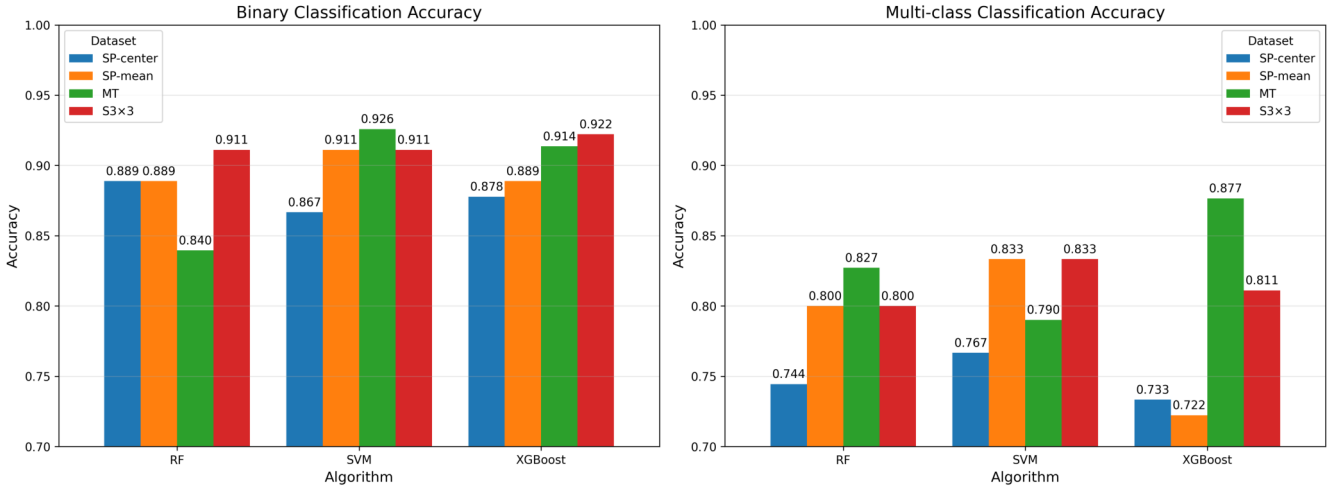


Fig. 2. Classification accuracy comparison across algorithms and dataset configurations

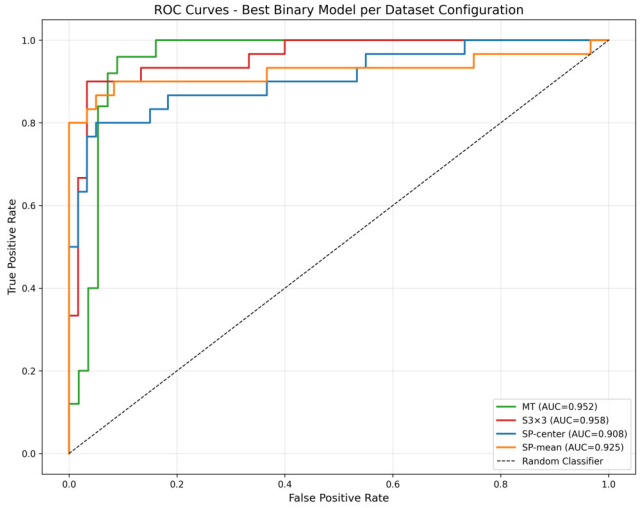


Fig. 3. ROC curves for binary hogweed detection

TABLE IX
XGBOOST PERFORMANCE ACROSS DATASET CONFIGURATIONS

Dataset	Task	Method	Accuracy
SP-center	Binary	Grid	0.878
SP-center	Multiclass	Default	0.733
SP-mean	Binary	Grid	0.889
SP-mean	Multiclass	Default	0.722
MT	Binary	Grid	0.914
MT	Multiclass	Grid	0.877
S3×3	Binary	Grid	0.922
S3×3	Multiclass	Default	0.811

D. Feature Importance Analysis

Figure 4 presents the top 15 most important features for the best-performing model (S3×3 configuration with XGBoost, 92.2% binary accuracy). Feature importance is computed based on the gain metric from XGBoost.

Spatial texture features, particularly standard deviation statistics, were prominent in the ranking, accounting for 7 of the top 15 features. The green band standard deviation (B03_std) emerged as the most important feature (0.222). Red-edge bands (B05, B06, B07) also appeared frequently across multiple statistical representations, consistent with their known sensitivity to vegetation biochemical properties.

The presence of built-up and bare soil indices (NDBI, BSI) among the top-ranked features suggests that contrast between hogweed and non-vegetated surfaces may contribute to the model's predictive performance within the local 3×3 neighborhood.

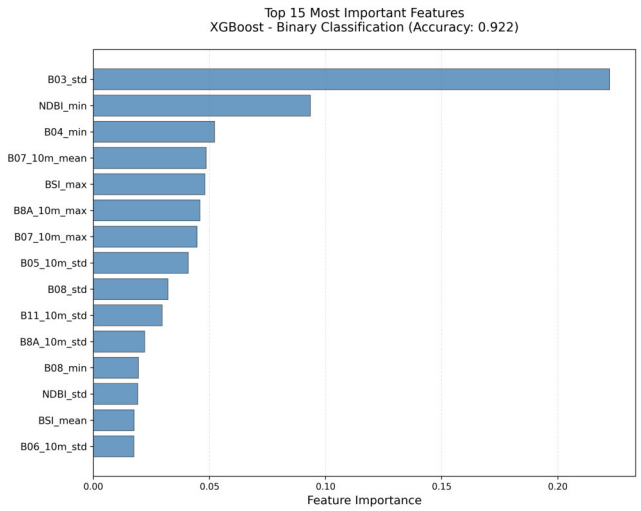


Fig. 4. Most important features for S3×3 XGBoost model (92.2% accuracy)

VI. CONCLUSIONS

This study demonstrates that the automated detection of Sosnowsky's hogweed (Heracleum sosnowskyi) using freely available Sentinel-2 multispectral imagery is a viable and scalable alternative to the current manual re-inspection process

maintained by the State Plant Protection Service. Among the evaluated satellite platforms, Sentinel-2 was identified as the most suitable for this task, owing to its spatial and temporal resolutions, freely accessible data archive, and the broad range of available spectral bands.

All three evaluated algorithms — SVM, Random Forest, and XGBoost — produced competitive results. However, SVM achieved the highest overall binary classification accuracy (92.6%, ROC-AUC 0.952 on the multi-temporal configuration), making it the recommended algorithm for this problem. Binary classification consistently outperformed multi-class classification across all dataset configurations, with accuracy gaps of up to 12.2%. For practical deployment, the binary classification formulation is therefore preferred.

Despite these promising results, the method was evaluated on a curated dataset of 450 reference samples. Before the framework can be considered operational for large-scale monitoring, it must be validated on broader geographic extents.

Future work should also investigate the applicability of the method for detecting previously unregistered hogweed occurrences, which would further extend its utility for national-scale invasive species management.

Beyond hogweed detection, the proposed framework has the potential to serve as a foundation for broader invasive species monitoring efforts. Should the method prove effective in real-world scenarios, the underlying pipeline could in principle be adapted for other invasive species present in Latvia, such as Canadian goldenrod (*Solidago canadensis*) or Himalayan balsam (*Impatiens glandulifera*).

VII. Acknowledgement

During the preparation of this work, the author(s) used Claude Sonnet 4.6 for translation and editorial correction. The author(s) take(s) full responsibility for the content of this publication.

VIII. References

References

[1] *Heracleum sosnowskyi*, Skatīts: 2026. gada 13. aprīlis, 03/2026.

[2] V. augu aizsardzības dienests, *Sosnovska latvānis*, Skatīts: 2026. gada 17. marts, 02/2022.

[3] V. augu aizsardzības dienests, *Sosnovska latvāņa izplatība latvijā*, Skatīts: 2026. gada 17. marts, 07/2025.

[4] J. Müllerová, J. Brůna, T. Bartaloš, P. Dvořák, M. Vítková, and P. Pyšek, "Timing is important: unmanned aircraft vs. satellite imagery in plant invasion monitoring", Frontiers in Plant Science **Volume 8 - 2017**, 10.3389/fpls.2017.00887 (2017).

[5] M. M. Zaka and A. Samat, "Advances in remote sensing and machine learning methods for invasive plants study: a comprehensive review", Remote Sensing **16**, 10.3390/rs16203781 (2024).

[6] W. Maes and K. Steppe, "Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture", Trends in Plant Science **24**, 10.1016/j.tplants.2018.11.007 (2018).

[7] J. Morgan, S. Gergel, and N. Coops, "Aerial photography: a rapidly evolving tool for ecological management", Bioscience **60**, 47–59 (2010).

[8] R. Massey, L. Bernier, A. Foster, S. Goetz, and U. Vepakomma, "Remote sensing tools for monitoring forests and tracking their dynamics", in (02/2023).

[9] Z. Sun, X. Wang, Z. Wang, L. Yang, Y. Xie, and Y. Huang, "Uavs as remote sensing platforms in plant ecology: review of applications and challenges", Journal of Plant Ecology **14**, 1003–1023 (2021).

[10] D. Olson and J. Anderson, "Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture", Agronomy Journal **113**, 971–992 (2021).

[11] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: a review", Remote Sensing **12**, 10.3390/rs12142291 (2020).

[12] D. Roy, M. Wulder, T. Loveland, C. Woodcock, R. Allen, M. Anderson, D. Helder, J. Irons, D. Johnson, R. Kennedy, T. Scambos, C. Schaaf, J. Schott, Y. Sheng, E. Vermote, A. Belward, R. Bindschadler, W. Wohen, F. Gao, and Z. Zhu, "Landsat-8: science and product vision for terrestrial global change research", Remote Sensing of Environment **2014**, 154–172 (2014).

[13] J. Masek, M. Wulder, B. Markham, J. McCorkel, C. Crawford, J. Storey, and D. Jenstrom, "Landsat 9: empowering open science and applications through continuity", Remote Sensing of Environment **248**, 111968 (2020).

[14] European Space Agency (ESA), *Sentinel-2 user handbook*, Issue 1, Rev. 2. Available online: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook, European Space Agency (Paris, France, 2015).

[15] O. Justice, E. Vermote, J. Townshend, R. Defries, D. Roy, D. Hall, V. Salomonson, J. Privette, G. Riggs, A. Strahler, W. Lucht, R. Myneni, Y. Knyazikhin, S. Running, R. Nemani, Z. Wan, A. Huete, W. Van Leeuwen, R. Wolfe, and M. Barnsley, "The moderate imaging spectroradiometer (modis): land remote sensing for global change research. ieee trans. geosci. remote sens", Geoscience and Remote Sensing, IEEE Transactions on **36**, 1228–1249 (1998).

[16] R. Houborg and M. McCabe, "Daily retrieval of ndvi and lai at 3 m resolution via the fusion of cubesat, landsat, and modis data", Remote Sensing **10**, 10.3390/rs10060890 (2018).

[17] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: planetary-scale geospatial analysis for everyone", Remote Sensing of Environment **202**, Big Remotely

Sensed Data: tools, applications and experiences, 18–27 (2017).

[18] Valsts Augu Aizsardzības Dienests, *AKD Latvāņi public shapes: hogweed distribution database*, ArcGIS REST Services, State Plant Protection Service of Latvia., 2025.

[19] SentiWiki, *Sentinel-2 products*, Copernicus Sentinel Online, 2023.

[20] World Bank, *Latvia: land use statistics*, World Bank Open Data, 2023.