TELEDETEKCJA ŚRODOWISKA
dawniej „Fotointerpretacja w geografii"
Tom 64 (2024) s. 24–38
ISSN 0071-8076
POLSKIE TOWARZYSTWO GEOGRAFICZNE
ODDZIAŁ TELEDETEKCJI
http://ptgeo.org.pl/teledetekcja/

Analysis of the Effectiveness of Selected Machine Learning Algorithms in the Classification of Satellite Image Content Depending on the Size of the Training Sample

Analiza skuteczności wybranych algorytmów uczenia maszynowego w klasyfikacji treści obrazów satelitarnych w zależności od rozmiaru próbki treningowej

Przemysław KUPIDURA
Warsaw University of Technology, Faculty of Geodesy and Cartography
przemyslaw.kupidura@pw.edu.pl

Stanisław NIEMYSKI
Warsaw University of Technology, Faculty of Geodesy and Cartography
sniemyski@o2.pl

Abstract: The article presents an analysis of the accuracy of 3 popular machine learning (ML) methods: Maximum Likelihood Classifier (MLC), Support Vector Machine (SVM), and Random Forest (RF) depending on the size of the training sample. The analysis involved performing the classification of the content of a Landsat 8 satellite image (divided into 6 basic land cover classes) in 10 different variants of the number of training samples (from 2664 to 34711 pixels), estimating individual results, and a comparative analysis of the obtained results. For each classification variant, an error matrix was developed and on their basis, accuracy metrics were calculated: f1-score, precision and recall (for individual classes) as well as overall accuracy and kappa index of agreement (generally for the entire classification). The analysis showed a stimulating effect of the size of the training sample on the accuracy of the obtained classification results in all analyzed cases, with the most sensitive to this factor being MLC, showing the best effectiveness with the largest training sample and the smallest - with the smallest, and the least SVM, characterized by the highest accuracy with the smallest training sample, comparing to other algorithms.

Streszczenie: Artykuł przedstawia analizę dokładności 3 popularnych metod uczenia maszynowego: Maximum Likelihood Classifier (MLC), Support Vector Machine (SVM) oraz Random Forest (RF) w zależności od liczebności próbki treningowej. Analiza polegała na wykonaniu klasyfikacji treści zdjęcia satelitarnego Landsat 8 (w podziale na 6 podstawowych klas pokrycia terenu) w 10 różnych wariantach liczebności próbek uczących (od 2664 do 34711 pikseli), oszacowaniu poszczególnych wyników oraz analizie porównawczej uzyskanych wyników. Dla każdego wariantu klasyfikacji opracowano macierz błędów, a na ich podstawie obliczono metryki dokładności: F1-score, precision and recall (dla pojedynczych klas) oraz ogólną dokładność i wskaźnik zgodności Kappa (ogólnie dla całej klasyfikacji). Analiza wykazała stymulujący wpływ rozmiaru próbki uczącej na dokładność uzyskiwanych wyników klasyfikacji we wszystkich analizowanych przypadkach, przy czym najbardziej wrażliwym na ten czynnik był MLC, wykazujący się najlepszą skutecznością przy największej próbce treningowej i najmniejszą – przy najmniejszej, a najmniej SVM, cechujący się największą dokładnością przy najmniejszej próbce treningowej, w porównaniu do pozostałych algorytmów.

Keywords: machine learning, classification; remote sensing; training sample size; SVM, Random Forest, Maximum Likelihood Classifier, satellite imagery

Słowa kluczowe: uczenie maszynowe; klasyfikacja; teledetekcja; rozmiar próbki treningowej, SVM, lasy losowe, klasyfikator największego prawdopodobieństwa; zobrazowania satelitarne

Introduction

The accurate classification of land use and land cover (LULC) is a critical task in remote sensing applications. Over the years, the development of machine learning (ML) algorithms has significantly advanced the (semi)automatic classification of aerial and satellite images. However, the abundance of available options poses challenges in selecting the optimal solution.

Numerous scientific articles have been dedicated to the comparison of various ML methods in the context of satellite image content classification. Notable works include (starting from the most recent): Bidgeli et al. (2024), Ding (2024), Zhao et al. (2024), Mousavinezhad et al. (2023), Seydi et al. (2023), Sobieraj et al. (2022), Ghayour et al. (2021), Liu et al. (2021), Koppaka & Moth (2020), Volke & Abarca-del-Rio (2020), Maxwell et al. (2018, 2015, 2014a, 2014b), Li et al. (2016), Maxwell & Warner (2015), Cracknell & Reading (2014), Duro et al. (2012), Bukholder et al. (2011). The aforementioned studies analyze selected algorithms from a long list of those popular in remote sensing, including: Minimum Distance (MD), Maximum Likelihood (MLC), Decision Trees (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). The results of these studies are not unequivocally conclusive (although it can be noted that RF and SVM often demonstrate very good effectiveness), as the effectiveness of individual methods may depend on the nature of the classification task, the processed data, and the size or quality of the training sample.

This very topic: the effectiveness of ML algorithms in classifying satellite image content depending on the size of the training sample - can also be found in the scientific literature on the subject. Works such as Shang et al. (2018), Ramezan et al. (2021), Fu et al. (2023), and Zheng et al. (2020) can be mentioned here. They clearly indicate a positive correlation between classification effectiveness and the size of the training sample, although this relationship varies depending on the algorithm. Similar conclusions are drawn from other studies, dedicated to other issues, not directly related to remote sensing (Budach et al. 2022; Figueroa et al. 2012; Halevy et al. 2009; Raudys et al. 1991).

The research presented in this article is another attempt to answer the question about the effectiveness of selected ML methods depending on the size of the training sample. Three popular algorithms were analyzed: MLC, RF, and SVM. The classification of the Landsat-8 image was carried out in 10 different variants - with different numbers of training samples, from 2664 to 34711 pixels. This allowed for the examination of the impact of the training sample size on the effectiveness of the three analyzed ML algorithms.

Materials and Methods

Analiza składała się z następujących etapów:

1. Vectorization of the test area on the satellite scene, divided into land cover classes.
2. Selection of 10 groups of training fields for each class - two variants of training data were prepared.
3. Classification of land cover according to the adopted taxonomy, using selected ML algorithms.
4. Validation of accuracy based on the entire vectorized test area.

The methodology is also presented in Fig. 1.

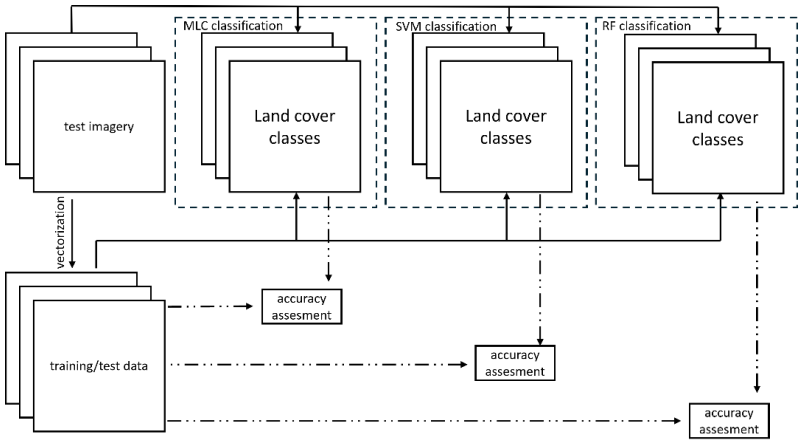

Fig.1. Research methodology scheme
Fig. 1. Schemat medotyki badania

Test image and test area

A Landsat-8 satellite scene from May 19, 2017, covering the area of northwestern Poland, was selected for the study (Fig. 2). Six spectral bands were used: 2 (blue), 3 (green), 4 (red), 5 (near infrared), 6 (shortwave infrared 1), and 7 (shortwave infrared 2). The image ground sample distance (GSD) is 30 m.

Fig.2. Test area and test image
Fig. 2. Obszar i obraz testowy

Land cover classes

Six basic land cover classes were distinguished. Since these are mostly classes composed of many different clusters, and thus do not show an ellipsoidal character in the feature space, subclasses were distinguished within these classes, which were subjected to proper classification, and the results were aggregated to these six basic classes.

The selection of training fields was carried out in two basic variants. In variant A, a large number of subclasses were identified, which fully exhausted the types of land cover within the basic classes. In variant B, several subclasses were deliberately omitted, in order to check the impact of the quality of the training sample on the accuracy of classification - by comparing the results obtained for variants A and B. The division into classes and subclasses is presented in Table 1. Subclasses defining the same type of land cover (e.g., Rapeseed 1 and Rapeseed 2) were distinguished based on the analysis of differences in spectral reflectance. The accuracy of the land cover maps obtained as a result of this aggregation/reclassification was evaluated.

Table 1. Set of classes and subclasses in the analyzed variants.
Tabela 1. zestaw klas i podklas w analizowanych wariantach.

Class name	Subclass name Variant A	Subclass name Variant B
Water	Deep water	Water
	Shallow water	
	Cloudy water	
Built-up area	Built-up area 1	Built-up area 1
	Built-up area 2	---
Soil	Dry soil	Dry soil
	Moderately moist soil	Moderately moist soil
	Moist soil	Moist soil
Low vegetation	Rapeseed 1	Rapeseed 1
	Rapeseed 2	---
	Cereal 1	---
	Cereal 2	Cereal 2
	Cereal 3	Cereal 3
	Permanent grassland 1	Permanent grassland 1
	Permanent grassland 2	---
	Permanent grassland 3	---
High vegetation	Coniferous forest	Coniferous forest
	Mixed forest	Mixed forest
	Deciduous forest 1	Deciduous forest 1
	Deciduous forest 2	---
	Bushes	Bushes
Wetland	Wetland	Wetland

Machine learning algorithms

Three ML algorithms were analyzed: MLC, SVM, and RF, briefly characterized below. Authors used ArcGIS software.

Maximum Likelihood

The model built on the basis of this algorithm is based on maximizing the probability function, so that within the assumed statistical model, elements of the set are assigned to classes to which their membership is most probable. It is indicated for its high effectiveness with a large amount of training data, with a distribution close to normal, of good quality (without noise/erroneous observations) (Kelley, 2021).

Support Vector Machine

The model built according to this algorithm is based on hyperplanes that maximize the distance between classes in a multidimensional feature space (Boser et al. 1992; Cortes & Vapnik, 1995). Originally developed for distinguishing classes with linear separation, the use of kernel functions also allows for solving nonlinear problems (Schölkopf & Smola, 2001). This method works well, among others, in issues related to a large number of features (Nalepa & Kawulok, 2019).

Random Forest

This is an example of ensemble learning. The Random Forest model is based on a certain number of decision trees, created independently based on bagging (bootstrap aggregation), which is the random selection of training samples, but also feature randomness. This allows for maintaining a low correlation between trees, and as a result, also avoids overfitting, which is often a characteristic of a single decision tree. (Ho, 1995, 1998; Breiman, 2001). This is a method resistant to disturbances in the quality of training data (Belgiu & Drăguț, 2016).

Training data

The experiment was conducted twice, for two different test areas, differing in training data and their quantity. For each of the areas, 10 sets of training fields were prepared. Based on them, a series of independent classifications were carried out - in such a way that in the first variant - with the smallest size of training data - it was performed based on 1 training set, adding another set in each subsequent one. Details regarding the total number of training data are presented in Table 2.

Table 2. Number of pixels for training sample sizes in both variants.
Tabela 2. Liczba pikseli używana dla poszczególnych rozmiarów próbek treningowych w obydwu wariantach.

Training sample size	Variant A training pixel number	Variant B training pixel number
1	3867	2664
2	7473	5163
3	11336	7624
4	14484	9744
5	18753	12566
6	22004	14825
7	25284	17163
8	28701	19670
9	31820	21898
10	34711	23989

Accuracy asessment

The accuracy assessment was based on the entire vectorized test area. The characteristics of the test area, divided into classes, are presented in Table 3.

Table 3. Number of pixels in the test data.
Tabela 3. Liczba pikseli w danych testowych.

Class	Size of validation data [pixels]	Size of validation data [%]
Water	27300	2,1
Built-up area	27814	2,1
Soil	193246	14,9

Low vegetation	495929	38,2
High vegetation	544250	41,9
Wetland	10853	0,8

Based on the comparison of the results of individual classifications with test data, error matrices were developed. Based on these, precision, recall (Powers, 2007), and F1-score (Hand et al., 2021) values were calculated for each class, as well as the Kappa index of agreement (Sim & Wright, 2005) and overall accuracy (Labatut & Cherifi, 2012) for general classifications. The following formulas were applied:

For precision and recall (Hand et al., 2021):

$$precision = \frac{TP}{TP+FP},$$

$$recall = \frac{TP}{TP+FN},$$

where: *TP* = True Positive, *FP* = False Positive and *FN* = False Negative.

For F1-score (Hand et al., 2021):

$$F1 = \frac{2*precision*recall}{precision+recall}.$$

For Kappa index (Sim, Wright, 2005):

$$kappa = \frac{P_o - P_c}{1-P_c},$$

where: P_o = observed agreement, P_c = chance agreement.

For overall accuracy:

$$OA = \frac{C_c}{C_o},$$

where: C_c = number of pixels classified correctly, C_o = total number of classified pixels.

Results and discussion

First, the results obtained for individual classes were analyzed, followed by the overall classification. The results have been presented in tables and diagrams - in the following subsections. Additionally, selected output images of the classification have been presented in the Appendix.

Water

The results for the water class are presented in Table 4 and Figure 3.

Table 4. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size (TSS) for the *water* class.
Tabela 4. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy woda.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	1	0,105	0,190	1	0,007	0,014	1	0,999	0,999	1	0,962	0,981	0,707	1	0,828	0,992	0,994	0,993
2	1	0,456	0,626	1	0,016	0,031	0,998	1	0,999	0,999	0,992	0,995	0,997	0,998	0,997	1	0,973	0,986
3	1	0,564	0,721	1	0,251	0,401	0,997	1	0,998	1	0,989	0,994	0,998	1	0,999	1	0,995	0,997
4	1	0,605	0,754	1	0,401	0,572	0,997	1	0,998	0,999	1	0,999	0,988	1	0,994	1	1	1
5	1	0,778	0,875	1	0,615	0,762	0,996	1	0,998	0,999	0,999	0,999	0,998	1	0,999	1	1	1
6	1	0,948	0,973	1	0,661	0,796	0,998	1	0,999	0,998	1	0,999	0,999	1	0,999	1	1	1
7	1	0,976	0,988	1	0,698	0,822	0,996	1	0,998	0,998	0,997	0,997	0,945	1	0,972	1	1	1
8	1	0,974	0,987	1	0,776	0,874	0,996	1	0,998	0,998	0,999	0,998	0,996	1	0,998	1	1	1
9	1	0,977	0,988	1	0,819	0,900	0,99	1	0,999	0,999	0,999	0,999	0,995	1	0,997	0,999	1	0,999
10	1	0,988	0,994	1	0,823	0,903	1	0,999	0,999	0,999	1	0,999	0,988	1	0,994	0,999	1	0,999

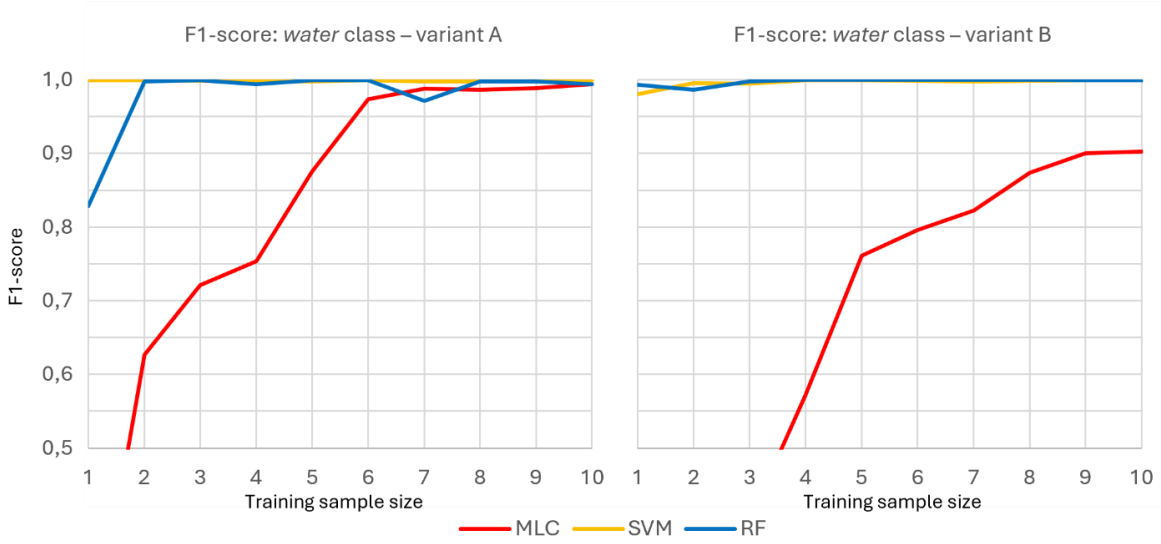

Fig. 3. Diagram of F1-score values for the *water* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 3. Wykres wartości F1-score dla klasy woda z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

The sensitivity of MLC to the size of the training sample is very noticeable. While SVM and RF are characterized by very good accuracy essentially for all sizes of the training sample, the results obtained using MLC are clearly worse, especially for training samples of smaller sizes. It is worth noting that this is due to very low recall values, and thus a large omission error. This may be related to an insufficiently representative distribution of training sample values for this class. It's worth noting that the *water* class is usually very well distinguishable from the other classes, due to its unique low spectral values, especially in the infrared range - this is shown by the very good results obtained for the SVM and RF methods. Significantly worse results for ML are probably due to an insufficiently large number of pixels allowing for statistically reliable calculation of the class signature. The values obtained for variant B are clearly worse for MLC than for variant A. This is probably not due to the size of the training sample itself, but to the fact of generalizing all subclasses of this class. As a result, the training sample stopped showing an ellipsoidal character, so the assumption of a normal distribution, effective with large training samples in variant A, gave an unsatisfactory result. In the case of using the other two tested methods, the analyzed values approach 1, in both variants, even with the smallest training samples.

Built-up area

The results for the built-up area class are presented in Table 5 and Figure 4.

Table 5. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size for the *built-up area* class.
Tabela 5. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy tereny zabudowane.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,136	0,996	0,239	0,144	0,993	0,252	0,889	0,948	0,918	0,893	0,677	0,770	0,790	0,923	0,851	0,698	0,660	0,678
2	0,320	0,992	0,484	0,280	0,988	0,436	0,951	0,877	0,913	0,913	0,873	0,893	0,921	0,890	0,905	0,874	0,788	0,829
3	0,366	0,990	0,534	0,324	0,981	0,487	0,962	0,913	0,937	0,953	0,705	0,810	0,917	0,921	0,919	0,915	0,805	0,856
4	0,475	0,989	0,642	0,434	0,987	0,603	0,961	0,910	0,935	0,955	0,721	0,822	0,920	0,909	0,914	0,922	0,793	0,853
5	0,609	0,980	0,751	0,564	0,975	0,715	0,969	0,887	0,926	0,942	0,730	0,823	0,928	0,883	0,905	0,859	0,700	0,771
6	0,699	0,979	0,816	0,596	0,972	0,739	0,980	0,877	0,926	0,970	0,740	0,840	0,953	0,875	0,912	0,894	0,658	0,758
7	0,703	0,977	0,818	0,601	0,972	0,743	0,974	0,865	0,916	0,975	0,748	0,84	0,931	0,862	0,895	0,862	0,714	0,781
8	0,733	0,978	0,838	0,646	0,976	0,777	0,968	0,886	0,925	0,962	0,747	0,841	0,903	0,867	0,885	0,943	0,689	0,796
9	0,770	0,973	0,860	0,683	0,975	0,803	0,964	0,878	0,919	0,966	0,778	0,862	0,960	0,889	0,923	0,931	0,720	0,812
10	0,788	0,971	0,870	0,696	0,973	0,812	0,779	0,976	0,866	0,979	0,766	0,860	0,964	0,876	0,918	0,961	0,701	0,811

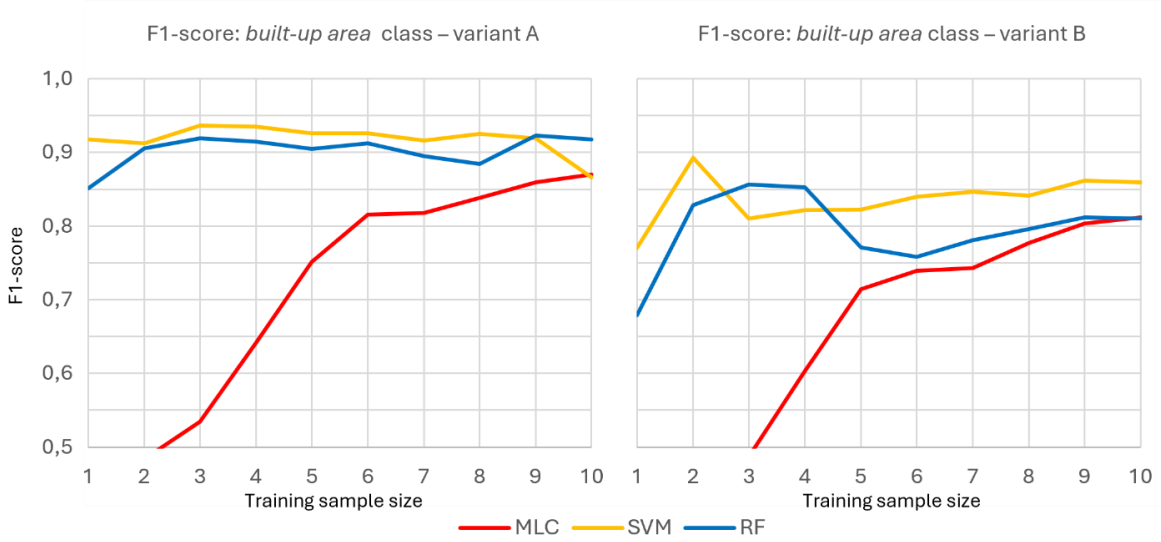

Fig. 4. Diagram of F1-score values for the *built-up area* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 4. Wykres wartości F1-score dla klasy tereny zabudowane z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

The accuracy obtained for this class is noticeably lower than for the water class, which results from the great diversity of the built-up area class and the lesser distinctiveness of its spectral values. The values obtained for SVM and RF are quite similar (although generally slightly better for SVM), some fluctuations in results may result from random errors, but a relatively constant level of accuracy can be observed, regardless of the size of the training sample. It looks different in the case of MLC, where, as with the water class, the accuracy of identification very clearly depends on the size of the training sample. Again, for samples of the smallest sizes, the values obtained are very low. This time it results from small *precision* values and is largely related to the water class: there is a large overestimation of built-up areas at the expense of water. The values for variant A are better than those obtained for variant B.

Soil

The results for the *soil* class are presented in Table 6 and Figure 5.

Table 6. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size for the *soil* class.
Tabela 6. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy gleba.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,995	0,759	0,861	0,994	0,873	0,930	0,999	0,971	0,985	0,942	0,999	0,970	0,971	0,999	0,985	0,938	0,999	0,938
2	0,987	0,949	0,968	0,986	0,951	0,968	0,972	0,999	0,985	0,966	1	0,983	0,968	0,999	0,983	0,939	1	0,969
3	0,991	0,961	0,976	0,989	0,962	0,975	0,975	0,999	0,987	0,946	0,999	0,972	0,975	0,998	0,986	0,953	0,999	0,975
4	0,989	0,953	0,971	0,988	0,953	0,970	0,976	0,999	0,987	0,950	0,999	0,974	0,971	0,999	0,985	0,952	0,999	0,975
5	0,978	0,983	0,980	0,979	0,982	0,980	0,967	0,999	0,983	0,948	0,999	0,973	0,962	0,998	0,980	0,936	1	0,967
6	0,979	0,982	0,980	0,979	0,981	0,980	0,970	0,998	0,984	0,952	0,999	0,975	0,962	0,998	0,980	0,934	1	0,966
7	0,980	0,978	0,979	0,981	0,978	0,979	0,966	0,999	0,982	0,951	0,999	0,974	0,959	0,997	0,978	0,939	1	0,969
8	0,982	0,977	0,979	0,982	0,977	0,979	0,968	0,999	0,983	0,951	0,999	0,974	0,961	0,991	0,976	0,938	1	0,968
9	0,983	0,977	0,980	0,983	0,977	0,980	0,971	0,999	0,985	0,956	0,999	0,977	0,968	0,999	0,983	0,943	1	0,971
10	0,976	0,976	0,976	0,946	0,977	0,976	0,998	0,955	0,976	0,949	0,999	0,973	0,964	0,999	0,981	0,938	1	0,968

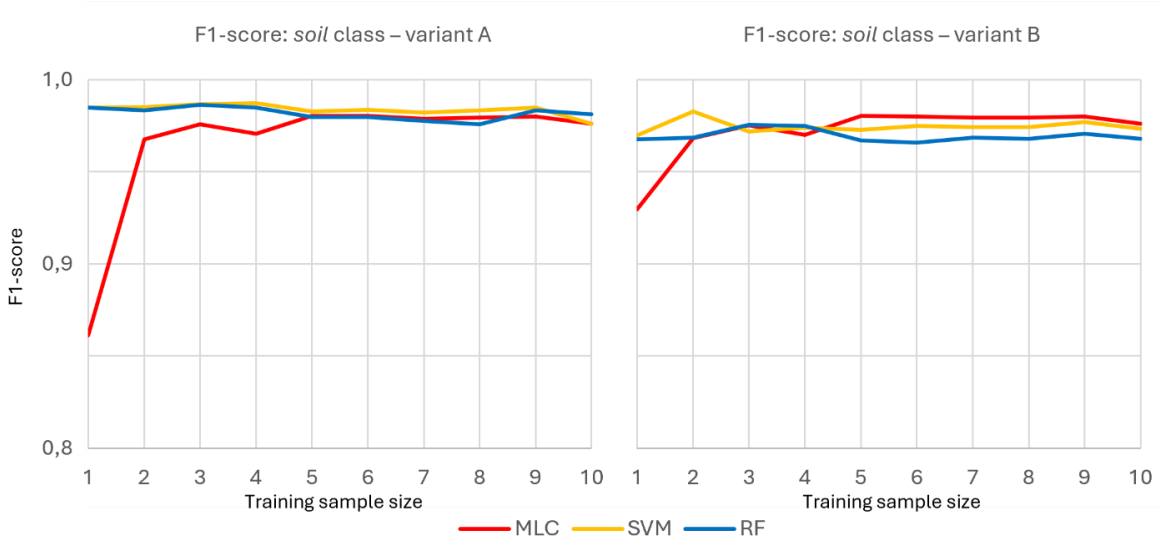

Fig. 5. Diagram of F1-score values for the *soil* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 5. Wykres wartości F1-score dla klasy gleba z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

The results obtained for the soil class are quite similar for all algorithms, excluding the effectiveness of MLC for the smallest training samples. In other cases, it is high accuracy with fairly similar values across the entire spectrum of tested training samples.

Low vegetation

The results for the *low vegetation* class are presented in Table 7 and Figure 6.

Table 7. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size for the *low vegetation* class.
Tabela 7. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy niska roślinność.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,967	0,885	0,924	0,982	0,727	0,835	0,895	0,955	0,924	0,953	0,922	0,937	0,949	0,881	0,914	0,942	0,871	0,905
2	0,944	0,967	0,955	0,976	0,899	0,936	0,951	0,898	0,924	0,944	0,935	0,939	0,948	0,942	0,945	0,945	0,879	0,911
3	0,951	0,929	0,940	0,973	0,845	0,904	0,955	0,896	0,925	0,946	0,928	0,937	0,951	0,901	0,925	0,943	0,895	0,918
4	0,950	0,961	0,955	0,970	0,910	0,939	0,955	0,894	0,923	0,943	0,925	0,934	0,951	0,905	0,927	0,937	0,871	0,903
5	0,950	0,969	0,959	0,943	0,967	0,955	0,955	0,906	0,930	0,934	0,939	0,936	0,944	0,932	0,938	0,952	0,908	0,929
6	0,949	0,971	0,960	0,941	0,980	0,960	0,955	0,907	0,930	0,936	0,940	0,938	0,952	0,943	0,947	0,952	0,943	0,947
7	0,947	0,975	0,961	0,939	0,986	0,962	0,953	0,920	0,936	0,945	0,927	0,936	0,949	0,956	0,952	0,947	0,946	0,946
8	0,946	0,982	0,964	0,940	0,986	0,962	0,954	0,920	0,937	0,943	0,931	0,937	0,949	0,956	0,952	0,943	0,948	0,945
9	0,944	0,984	0,964	0,943	0,981	0,962	0,953	0,922	0,937	0,949	0,947	0,348	0,942	0,961	0,951	0,942	0,943	0,942
10	0,944	0,986	0,965	0,941	0,985	0,962	0,923	0,953	0,941	0,950	0,922	0,936	0,946	0,962	0,956	0,942	0,947	0,944

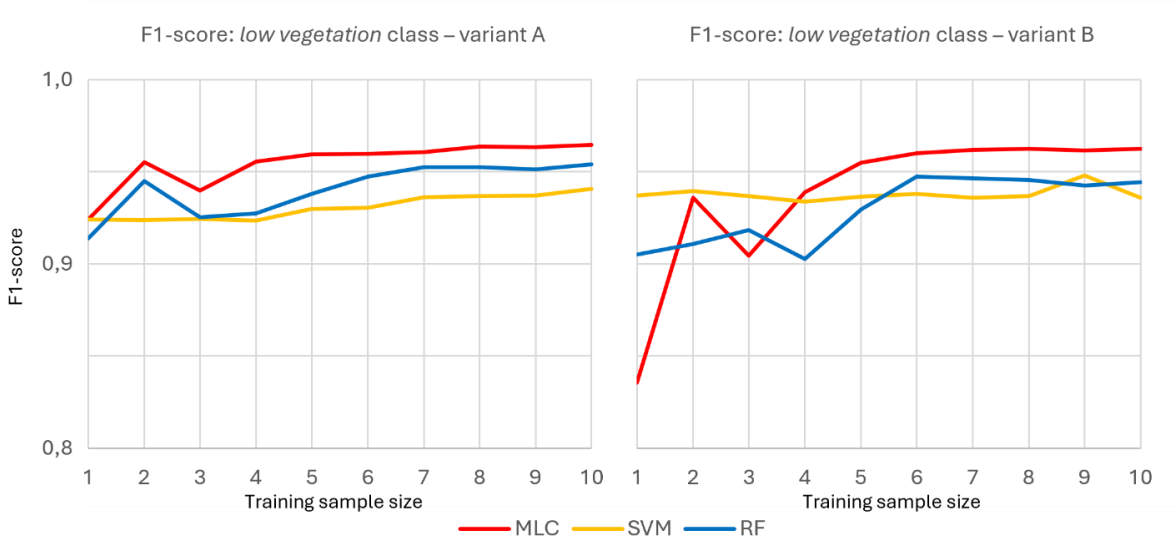

Fig. 6. Diagram of F1-score values for the *low vegetation* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 6. Wykres wartości F1-score dla klasy niska roślinność z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

We can observe a clear trend of increasing accuracy with the size of the training sample, although this trend varies depending on the algorithm. It is least visible in the case of SVM, where the values increase slightly, and strongest in the case of MLC. As a result, it can be noticed that for samples of the smallest sizes, SVM achieves the best effectiveness, while MLC - the worst (especially in the case of variant B), and for samples of larger sizes - the opposite: MLC gives the best results and SVM - the worst.

High vegetation

The results for the *high vegetation* class are presented in Table 8 and Figure 7.

Table 8. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size for the *high vegetation* class.
Tabela 8. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy wysoka roślinność.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,890	0,797	0,841	0,781	0,812	0,796	0,962	0,908	0,934	0,929	0,958	0,943	0,894	0,848	0,870	0,885	0,857	0,871
2	0,971	0,888	0,928	0,913	0,920	0,916	0,910	0,960	0,934	0,941	0,949	0,945	0,947	0,936	0,941	0,896	0,887	0,891
3	0,936	0,918	0,927	0,871	0,939	0,904	0,908	0,963	0,935	0,933	0,951	0,942	0,913	0,949	0,931	0,905	0,903	0,904
4	0,966	0,947	0,956	0,924	0,968	0,945	0,906	0,963	0,934	0,929	0,949	0,939	0,916	0,94	0,930	0,888	0,938	0,912
5	0,973	0,948	0,960	0,970	0,943	0,956	0,916	0,961	0,938	0,942	0,939	0,940	0,941	0,950	0,945	0,921	0,950	0,935
6	0,975	0,947	0,961	0,981	0,940	0,960	0,916	0,962	0,938	0,942	0,942	0,942	0,950	0,948	0,949	0,949	0,952	0,950
7	0,979	0,944	0,961	0,988	0,938	0,962	0,927	0,959	0,943	0,932	0,951	0,941	0,961	0,948	0,954	0,952	0,937	0,944
8	0,984	0,946	0,965	0,987	0,941	0,963	0,928	0,960	0,944	0,936	0,950	0,943	0,961	0,953	0,957	0,953	0,947	0,950
9	0,985	0,946	0,965	0,982	0,945	0,963	0,929	0,960	0,944	0,949	0,954	0,951	0,964	0,946	0,955	0,948	0,946	0,947
10	0,987	0,945	0,966	0,981	0,943	0,964	0,960	0,935	0,947	0,928	0,956	0,942	0,965	0,948	0,956	0,952	0,945	0,948

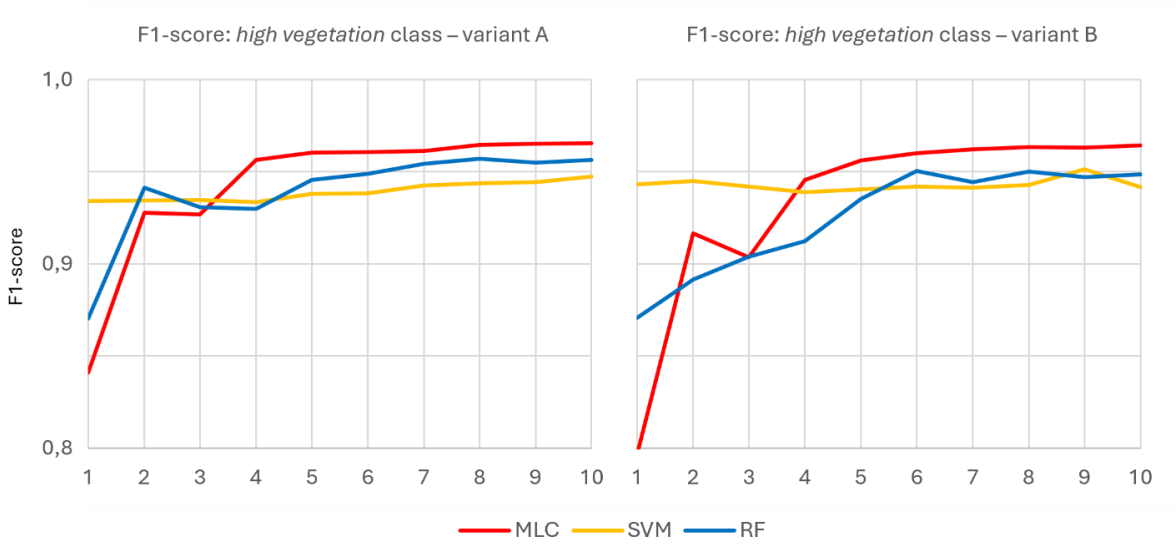

Fig. 7. Diagram of F1-score values for the *high vegetation* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 7. Wykres wartości F1-score dla klasy wysoka roślinność z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

The results obtained for the *high vegetation* class show a similar trend to the *low vegetation* class. The difference is that the dependence of the effectiveness of identifying this class on the size of the training sample is more noticeable, but only in the case of MLC and RF algorithms, which are clearly weaker than SVM for training samples of small sizes. However, the trend shown by SVM is almost identical to that of the low vegetation class.

Wetland

The results for the *wetland* class are presented in Table 9 and Figure 8.

Table 9. Values of precision (P), recall (R), and F1-score (F1) depending on the ML algorithm, variant, and training sample size for the *wetland* class.
Tabela 9. Wartości precision (P), recall (R) i F1-score (F1) w zależności od algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla klasy tereny podmokłe.

TSS	MLC						SVM						RF					
	Variant A			Variant B			Variant A			Variant B			Variant A			Variant B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0,971	0,370	0,536	0,956	0,398	0,562	0,997	0,528	0,690	0,999	0,529	0,692	0,078	0,382	0,130	0,070	0,350	0,117
2	0,234	0,198	0,215	0,234	0,198	0,215	0,998	0,519	0,683	0,997	0,477	0,645	0,350	0,487	0,407	0,130	0,504	0,207
3	0,854	0,246	0,382	0,854	0,246	0,382	0,995	0,517	0,680	0,996	0,536	0,697	0,492	0,478	0,485	0,169	0,484	0,251
4	0,999	0,305	0,467	0,999	0,305	0,467	0,998	0,533	0,695	0,998	0,532	0,694	0,424	0,783	0,452	0,552	0,478	0,512
5	0,999	0,315	0,479	0,999	0,315	0,479	0,999	0,490	0,658	0,998	0,478	0,646	0,957	0,547	0,696	0,691	0,543	0,608
6	0,999	0,430	0,601	0,999	0,430	0,601	0,996	0,542	0,702	0,998	0,535	0,697	0,557	0,582	0,569	0,805	0,555	0,657
7	0,996	0,469	0,638	0,996	0,469	0,638	0,997	0,504	0,670	0,995	0,542	0,702	0,826	0,535	0,469	0,499	0,566	0,530
8	0,992	0,483	0,650	0,992	0,483	0,650	0,998	0,511	0,676	0,992	0,530	0,691	0,878	0,553	0,679	0,918	0,566	0,700
9	0,993	0,550	0,708	0,993	0,550	0,708	0,999	0,522	0,686	0,998	0,551	0,710	0,916	0,565	0,699	0,905	0,550	0,684
10	0,992	0,552	0,709	0,992	0,552	0,709	0,998	0,524	0,687	0,996	0,480	0,648	0,841	0,551	0,666	0,851	0,545	0,664

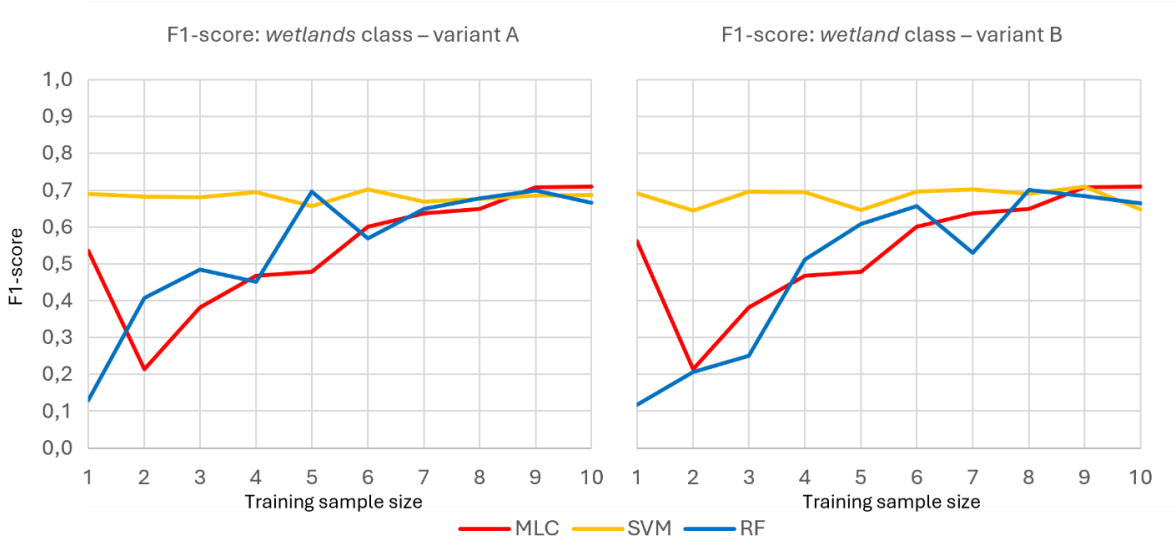

Fig. 8. Diagram of F1-score values for the *wetland* class using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 8. Wykres wartości F1-score dla klasy woda z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

Also in the case of the wetland class, it can be noticed that SVM shows relatively good accuracy with small sizes of the training sample, clearly better than RF and MLC. The difference between the results for this class and the results for the low and high vegetation classes is that for samples of the largest sizes, the accuracy of SVM is similar to MLC and RF, while for the vegetation classes it was noticeably lower.

Overall classification

The results for the entire classification are presented in Table 10 and Figure 9.

Table 10. Values of overall accuracy (OA) and kappa agreement coefficient (k) depending on the ML algorithm, variant, and training sample size for the entire image.
Tabela 10. Wartości ogólnej dokładności (OA) oraz współczynnika zgodności kappa (k) w zależności algorytmu uczenia maszynowego, wariantu i rozmiaru próbki treningowej dla całego obrazu.

TSS	MLC				SVM				RF			
	Variant A		Variant B		Variant A		Variant B		Variant A		Variant B	
	OA	k	OA	k	OA	K	OA	k	OA	k	OA	k
1	81,1	0,726	77,2	0,668	93,8	0,905	94,1	0,910	88,4	0,828	87,8	0,818
2	91,4	0,875	89,3	0,840	93,7	0,904	94,7	0,919	94,4	0,925	89,7	0,845
3	91,7	0,875	88,7	0,829	93,8	0,906	94,2	0,911	93,4	0,900	91,1	0,865
4	94,2	0,912	92,7	0,888	93,8	0,905	94,0	0,908	93,4	0,899	91,6	0,872
5	95,3	0,928	94,7	0,919	94,1	0,910	94,1	0,910	94,6	0,918	93,4	0,899
6	95,8	0,936	95,2	0,927	94,2	0,911	94,3	0,913	95,0	0,924	94,7	0,919
7	95,9	0,937	95,4	0,930	94,5	0,916	94,2	0,912	95,4	0,930	94,3	0,913
8	96,2	0,942	95,7	0,934	94,6	0,918	94,3	0,913	95,6	0,932	94,8	0,920
9	96,3	0,944	95,8	0,936	94,7	0,918	95,2	0,927	95,6	0,934	94,6	0,917
10	96,3	0,944	95,9	0,937	94,7	0,919	94,2	0,911	95,7	0,935	94,7	0,919

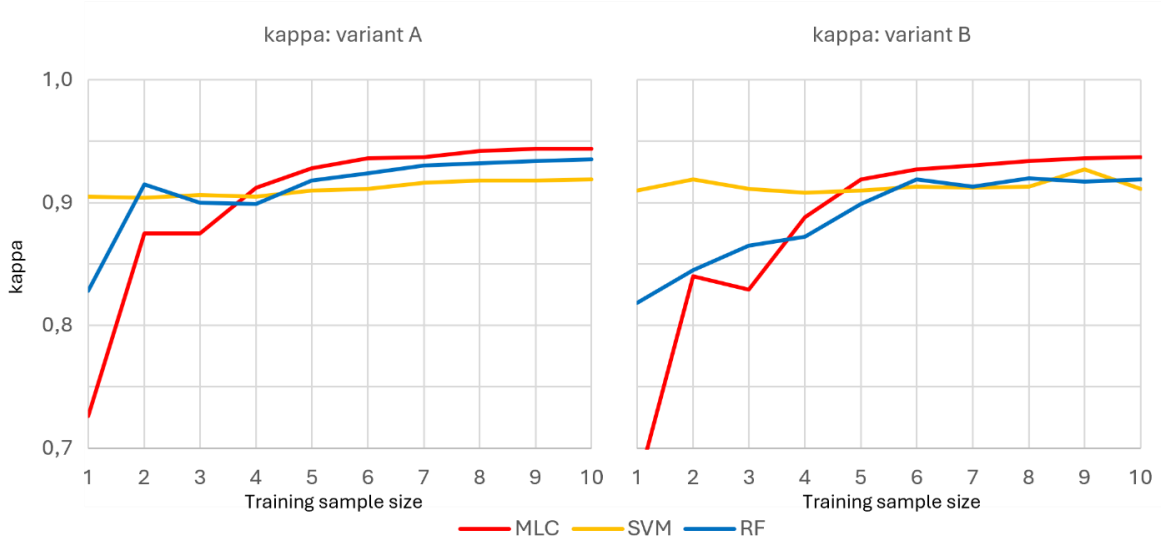

Fig. 9. Diagram of kappa values for the entire classification using individual ML algorithms depending on the training sample size, for variants A and B.
Ryc. 9. Wykres wartości kappa dla klasy całej klasyfikacji z użyciem poszczególnych algorytmów uczenia maszynowego w zależności od rozmiaru próbki treningowej, dla wariantów A i B.

Obviously, the results obtained for the entire classification are the consequence of the results for individual classes. Therefore, we can point out the relationships observed earlier - although for both variants they are similar, except that for variant A, generally better results were obtained. First of all, the accuracy generally increases with the size of the training sample. This trend is most noticeable in the case of MLC: for samples of the smallest sizes, the method gives very poor results, clearly worse than the best in such cases SVM (the difference in *Kappa* values exceeds 0.2). As TSS increases, the accuracy increases very significantly and is already the highest among the compared ones for samples of size 5 and larger. As a result, the difference between the highest and lowest accuracy for this method, among the tested training sample sizes, is - within a single variant - even 18 percentage points *OA* or over 0.2 *Kappa*. Also, a clearly upward trend with increasing TSS can be seen in the case of RF, although the differences between the best and worst result are not as large as in the case of MLC: approx. 8 percentage points *OA* or approx. 0.1 *Kappa*. TSS seems to have little effect on the effectiveness of SVM - the differences between the best and worst results obtained by this method are approx. 1 percentage point *OA* or less than 0.02 *Kappa*. It is worth noting here that for the smallest TSS, SVM gives clearly the best results, but at the largest - the worst.

Summary and conclusions

As part of the research, 60 scenarios of Landsat-8 satellite image classification were tested - for various ML algorithms: MLC, SVM, and RF, and different training sample sizes.

The research showed that MLC is characterized by the highest sensitivity to TSS, which gives very low classification effectiveness for small training samples, while for very large samples it surpasses the other analyzed methods in this respect. The least sensitive to TSS is SVM, which shows high result stability, regardless of TSS, with the smallest TSS being clearly the best against the other two algorithms, while with the largest TSS - slightly worse than them. RF, on the other hand, is in the middle of these methods: the impact of TSS on classification effectiveness is noticeable, but not as strong as in the case of MLC.

This indicates that while MLC is potentially a very effective algorithm, it is very sensitive to imperfections in the training data. It is worth noting that in most cases, classes were divided into coherent subclasses of an ellipsoidal nature in a distribution similar to normal, but when this division was abandoned, MLC showed very poor accuracy (especially compared to other algorithms), even with very large TSS. This proves that MLC requires very careful preparation of training data, numerous, noise-free and divided into ellipsoidal subclasses, regardless of the final classification systematics. On the other hand, SVM shows a very high resistance to imperfections in the training sample, giving good results also with a very small training sample and for non-ellipsoidal classes. It therefore appears as a "safe" choice, especially when there is no possibility of using well-prepared training data. By the way, the research results, especially those obtained for SVM, suggest the direction of further potential research: analysis of the accuracy of selected algorithms (mainly SVM) for even smaller TSS, in order to find the sample size for which a clear deterioration of SVM accuracy can be observed.

Author Contributions: Conceptualization, P.K. and S.N.; methodology, P.K. and S.N.; formal analysis, P.K.; investigation, S.N.; resources, S.N.; data curation, S.N.; writing, P.K..; visualization, P.K.; All authors have read and agreed to the published version of the manuscript."

Appendix A

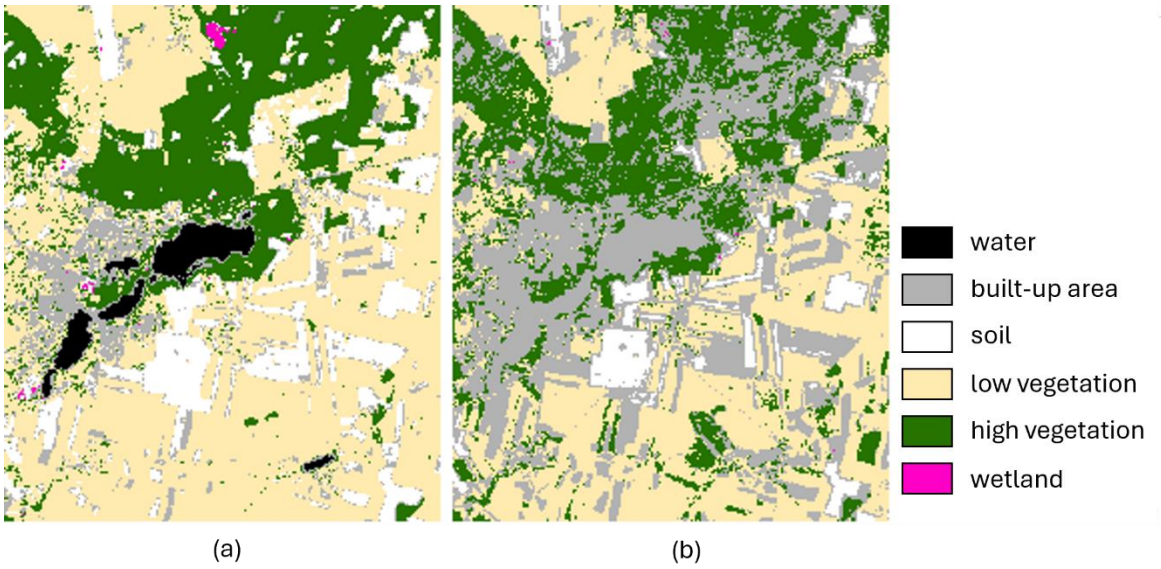

Fig. 10. Fragment of the image (10 km x 8.5 km) – the result of classification using MLC with the use of the maximum (a) and minimum (b) set of training data in variant A.
Ryc. 10. Fragment obrazu – wyniku klasyfikacji z wykorzystaniem MLC przy zastosowaniu maksymalnego (a) i minimalnego (b) zestawu danych treningowych w wariancie A.

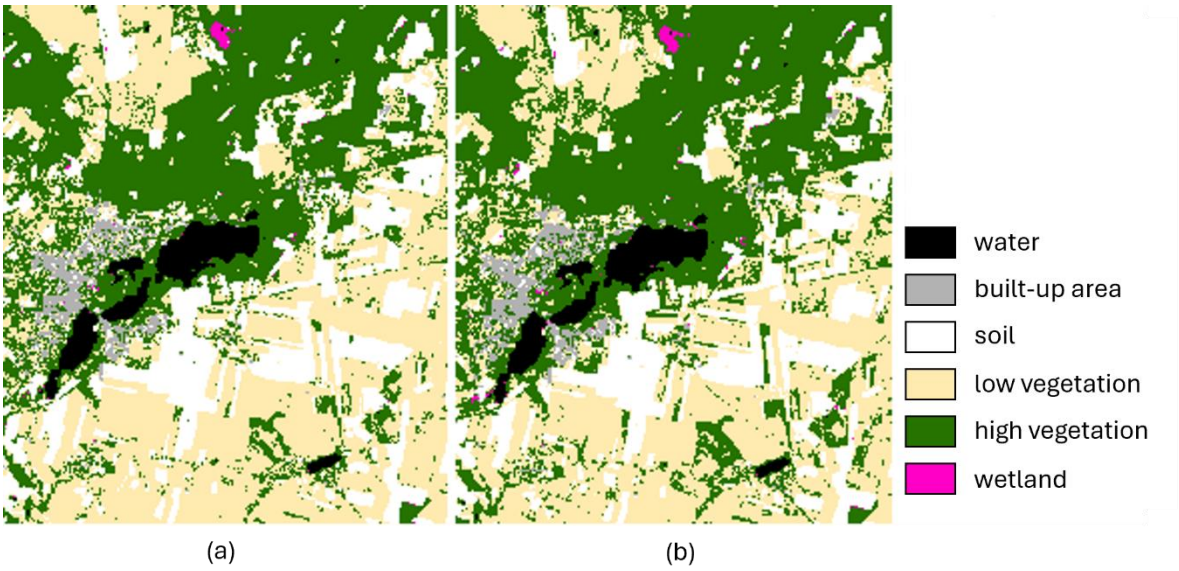

Fig. 11. Fragment of the image (10 km x 8.5 km) – the result of classification using SVM with the use of the maximum (a) and minimum (b) set of training data in variant A.
Ryc. 11. Fragment obrazu – wyniku klasyfikacji z wykorzystaniem SVM przy zastosowaniu maksymalnego (a) i minimalnego (b) zestawu danych treningowych w wariancie A.

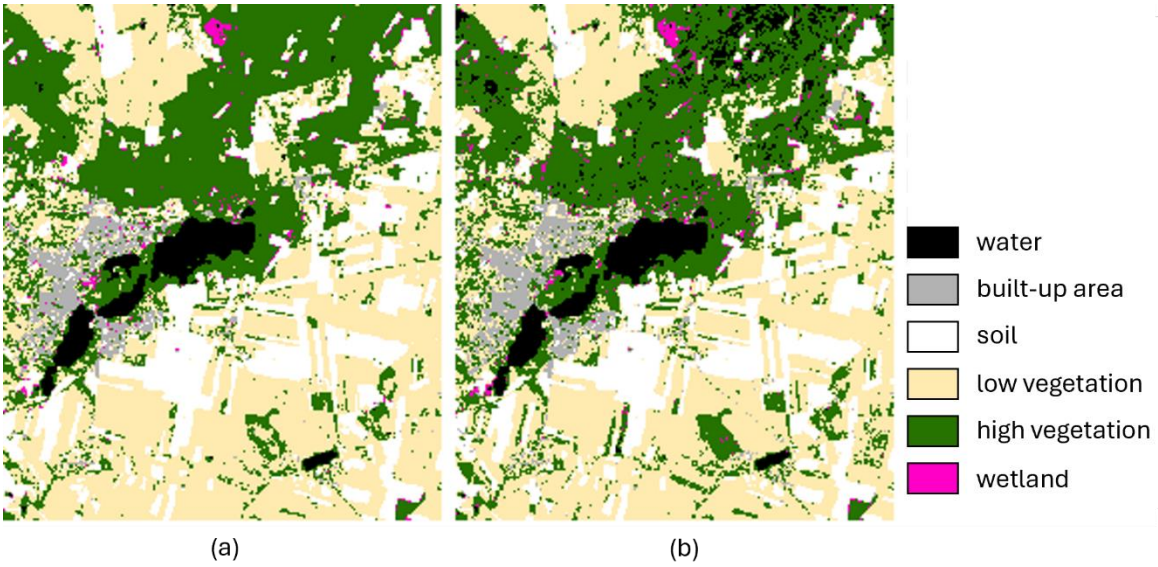

Fig. 12. Fragment of the image (10 km x 8.5 km) – the result of classification using RF with the use of the maximum (a) and minimum (b) set of training data in variant A.
Ryc. 12. Fragment obrazu – wyniku klasyfikacji z wykorzystaniem RF przy zastosowaniu maksymalnego (a) i minimalnego (b) zestawu danych treningowych w wariancie A.

References

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31. DOI: 10.1016/j.isprsjprs.2016.01.011

Bigdeli, A., Maghsoudi, A., & Ghezelbash, R. (2024). A comparative study of the XGBoost ensemble learning and multilayer perceptron in mineral prospectivity modeling: a case study of the Torud-Chahshirin belt, NE Iran. Earth Sci Inform, 17, 483–499. DOI: 10.1007/s12145-023-01184-4.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory – COLT '92 (p. 144). DOI: 10.1145/130385.130401

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. DOI: 10.1023/A:1010933404324

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022). The Effects of Data Quality on Machine Learning Performance. arXiv preprint arXiv:2207.14529.

Burkholder, A., Warner, T. A., Culp, M., & Landenberger, R. E. (2011). Seasonal trends in separability of leaf reflectance spectra for Ailanthus altissima and four other tree species. Photogramm. Eng. Remote Sens., 77, 793–804.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. DOI: 10.1007/BF00994018

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci., 63, 22–33.

Ding, H. (2024). Establishing a soil carbon flux monitoring system based on support vector machine and XGBoost. Soft Comput, 28, 4551–4574. DOI: 10.1007/s00500-024-09641-y.

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens. Environ., 118, 259–272.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., et al. (2012). Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making, 12(8). DOI: 10.1186/1472-6947-12-8

Fu, Y., Shen, R., Song, C., Dong, J., Han, W., Ye, T., & Yuan, W. (2023). Exploring the effects of training samples on the accuracy of crop mapping with machine learning algorithm. Science of Remote Sensing, Volume 7, 100081. DOI: 10.1016/j.srs.2023.100081

Ghayour, L., Neshat, A., Paryani, S., Shahabi, H., Shirzadi, A., Chen, W., Al-Ansari, N., Geertsema, M., Amiri, M. P., Gholamnia, M., et al. (2021). Performance Evaluation of Sentinel-2 and Landsat 8 OLI Data for Land Cover/Use Classification Using a Comparison between Machine Learning Algorithms. Remote Sensing, 13(7), 1349. DOI: 10.3390/rs13071349.

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2). DOI: 10.1109/MIS.2009.36

Hand, D. J., Christen, P., & Kirielle, N. (2021). F*: an interpretable transformation of the F-measure. Mach Learn, 110, 451–456. DOI: 10.1007/s10994-021-05964-1.

Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Kelley, P.R. (2021). Maximum Likelihood Estimation. In Foscher, M.M., Nijakamp, P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-60723-7_88

Koppaka, R., & Moh, T. -S. (2020). Machine Learning in Indian Crop Classification of Temporal Multi-Spectral Satellite Image. In 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). Taichung, Taiwan. DOI: 10.1109/IMCOM48794.2020.9001718.

Labatut, V., & Cherifi, H. (2012). Accuracy Measures for the Comparison of Classifiers. Proceedings of The 5th International Conference on Information Technology, Amman, Jordanie. 10.48550/arXiv.1207.3790.

Li, X., Chen, W., Cheng, X., & Wang, L. (2016). A Comparison of Machine Learning Algorithms for Mapping of Complex Surface-Mined and Agricultural Landscapes Using ZiYuan-3 Stereo Satellite Imagery. Remote Sensing, 8(6), 514. DOI: 10.3390/rs8060514.

Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., Zhang, J., Sun, Y., Guo, Z., Guo, Y., et al. (2021). Comparative Analysis of Two Machine Learning Algorithms in Predicting Site-Level Net Ecosystem Exchange in Major Biomes. Remote Sensing, 13(12), 2242. DOI: 10.3390/rs13122242.

Maxwell, A. E., & Warner, T. A. (2015). Differentiating mine-reclaimed grasslands from spectrally similar land cover using terrain variables and object-based machine learning classification. Int. J. Remote Sens., 36, 4384–4410.

Maxwell, A. E., Strager, M. P., Warner, T. A., Zegre, N. P., & Yuill, C. B. (2014). Comparison of NAIP orthophotography and RapidEye satellite imagery for mapping of mining and mine reclamation. GIScience Remote Sens., 51, 301–320.

Maxwell, A. E., Warner, T. A., Strager, M. P., Conley, J. F., & Sharp, A. L. (2015). Assessing machine-learning algorithms and image- and Lidar-derived variables for GEOBIA classification of mining and mine reclamation. Int. J. Remote Sens., 36, 954–978.

Maxwell, A. E., Warner, T. A., Strager, M. P., & Pal, M. (2014). Combining RapidEye satellite imagery and Lidar for mapping of mining and mine reclamation. Photogramm. Eng. Remote Sens.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. International Journal of Remote Sensing, 39(9), 2784-2817. DOI: 10.1080/01431161.2018.1433343

Mousavinezhad, M., Feizi, A., & Aalipour, M. (2023). Performance Evaluation of Machine Learning Algorithms in Change Detection and Change Prediction of a Watershed's Land Use and Land Cover. Int J Environ Res, 17, 29. DOI: 10.1007/s41742-023-00518-w

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. Artificial Intelligence Review, 52, 857–900. DOI: 10.1007/s10462-017-9611-1

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, Flinders University, Adelaide, Australia.

Ramezan, C.A., Warner, T.A., Maxwell, A.E., & Price, B.S. (2021). Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. Remote Sensing, 13(3), 368. DOI: 10.3390/rs13030368

Raudys, S. J., & Jain, A. K. (1991). Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(3).

Schölkopf, B., & Smola, A. J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press. ISBN 9780262256933

Seydi, S. T., Kanani-Sadat, Y., Hasanlou, M., Sahraei, R., Chanussot, J., & Amani, M. (2023). Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping. Remote Sensing, 15(1), 192. DOI: 10.3390/rs15010192

Shang, M., Wang, S.X., Zhou, Y. et al. (2018). Effects of Training Samples and Classifiers on Classification of Landsat-8 Imagery. Journal of the Indian Society of Remote Sensing, 46, 1333–1340. DOI: 10.1007/s12524-018-0777-z

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, 85(3), 257–268. DOI: 10.1093/ptj/85.3.257.

Sobieraj, J., Fernández, M., & Metelski, D. (2022). A Comparison of Different Machine Learning Algorithms in the Classification of Impervious Surfaces: Case Study of the Housing Estate Fort Bema in Warsaw (Poland). Buildings, 12(12), 2115. DOI: 10.3390/buildings12122115.

Volke, M. I., & Abarca-Del-Rio, R. (n.d.). Comparison of machine learning classification algorithms for land cover change in a coastal area affected by the 2010 Earthquake and Tsunami in Chile. Nat. Hazards Earth Syst. Sci. Discuss. [preprint].

Zhao, Z., Islam, F., Waseem, L. A., Tariq, A., Nawaz, M., Islam, I. U., Bibi, T., Rehman, N. U., Ahmad, W., Aslam, R. W., Raza, D., & Hatamleh, W. A. (2024). Comparison of Three Machine Learning Algorithms Using Google Earth Engine for Land Use Land Cover Classification. Rangeland Ecology & Management, 92, 129-137. https://doi.org/10.1016/j.rama.2023.10.007.

Zheng, W., & Jin, M. (2020). The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. SN Computer Science, 1, 71. DOI: 10.1007/s42979-020-0074-0

Przemysław Kupidura
Stanisław Niemyski
Polish Geographical Society
Remote Sensing Branch – Warszawa 2024