UDC 004.2
DOI: 10.20535/SRIT.2308-8893.2024.1.07

STATISTICAL METHODS OF FEATURE ENGINEERING FOR THE PROBLEM OF FOREST STATE CLASSIFICATION USING SATELLITE DATA

Y.V. SALII, A.M. LAVRENIUK, N.M. KUSSUL

Abstract. Timely detection of forest diseases is an important task for their prevention and spread limitation. The usage of satellite imagery provides capabilities for large-scale forest monitoring. Machine learning models allow to automate the analysis of these data for anomaly detection indicating diseases. However, selecting informative features is key to building an effective model. In this work, the application of Bhattacharyya distance and Spearman's rank correlation coefficient for feature selection from satellite images was investigated. A greedy algorithm was applied to form a subset of weakly correlated features. The experiment showed that selected features allow for improving the classification quality compared to using all spectral bands. The proposed approach demonstrates effectiveness for informative and weakly correlated feature selection and can be utilized in other remote sensing tasks.

Keywords: Sentinel-2, vegetation indices, Bhattacharyya distance, feature engineering, greedy algorithms, Spearman's rank correlation coefficient.

INTRODUCTION

Monitoring the condition of forest areas is a task for successfully identifying tree diseases and preventing their further spread. The use of high-resolution satellite images makes it possible to regularly obtain up-to-date information on large forest areas [1]. The process of processing and analyzing these data can be automated using machine learning methods that can detect signs of abnormal vegetation changes that may indicate the presence of diseases [2; 3].

One of the key steps in building an effective machine learning model for classification of the forest condition is the careful selection of the most informative features of the input data for the machine learning model. This allows to simplify the model and reduce the training time, without losing the quality of the classification. There are a large number of approaches to evaluating the informativeness of features. Most approaches are statistical, based on evaluating the similarity of data distributions. Among the most well-known methods for assessing the similarity of distributions, it is worth noting the Kullback–Leibler divergence, which calculates the relative entropy between two probability distributions. The higher the divergence value, the more distinct the distributions are [4]. However, this characteristic is not symmetrical, which limits its use. More universal are methods for determining the distance between data distributions, which include the Euclidean metric, that calculates the Euclidean distance between the means of two distributions, the Wasserstein distance, which measures the minimum "work" required to transform one distribution into another, or the chi-square distance, that compares the frequencies of samples from two distributions. Another symmetric

metric that allows you to measure the difference between two distributions is the Bhattacharyya distance [5].

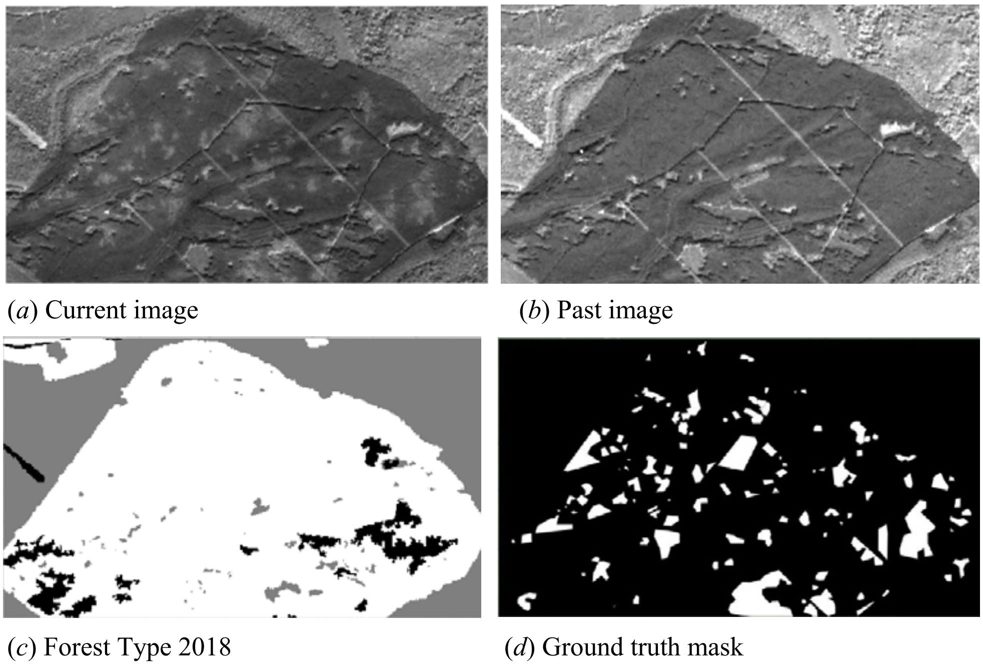
This paper investigates the possibility of using the Bhattacharyya distance and the Spearman correlation coefficient to select the most informative and at the same time weakly correlated features of multispectral satellite images.

FORMULATION OF THE PROBLEM

Let us consider the task of detecting disease in forest areas based on the analysis of Sentinel-2 satellite images [6]. The goal of our research is to develop an effective model that will be able to automatically determine whether a certain area of the forest is diseased on the basis of multispectral satellite data at different times.

To achieve this goal, two multispectral images will be used: current (Fig. 1, *a*) and past (Fig. 1, *b*). Since coniferous forests were studied, past images are not limited to a specific date, but it is important that the forest is healthy. Additionally, a vector mask of forest type (Fig. 1, *c*) from the Forest Type 2018 geospatial dataset [6] is used, which allows us to determine the areas where the forest is located and exclude non-forest areas from the analysis.

In the terminology of machine learning, the task is to build a binary classifier of each pixel of a multispectral image into the classes “healthy” and “stressed” (Fig. 1, *d*) by building and training a machine learning model. For training and testing of the model, the experts provided a ground truth (disease) mask (Fig. 1, *d*), which will be used for training the model and evaluation of its effectiveness. The experimental study was conducted on the data of the eastern part of the Grand Est region, France. The area for which training data is available is shown in Fig. 2.



(*a*) Current image

(*b*) Past image

(*c*) Forest Type 2018

(*d*) Ground truth mask

Fig. 1. Example of input data: (*a*, *b*) RGB (B4, B3, B2) composite of Sentinel-2 images; (c) white — coniferous, gray — deciduous, black — non-forest; (*d*) white — sick, black — healthy

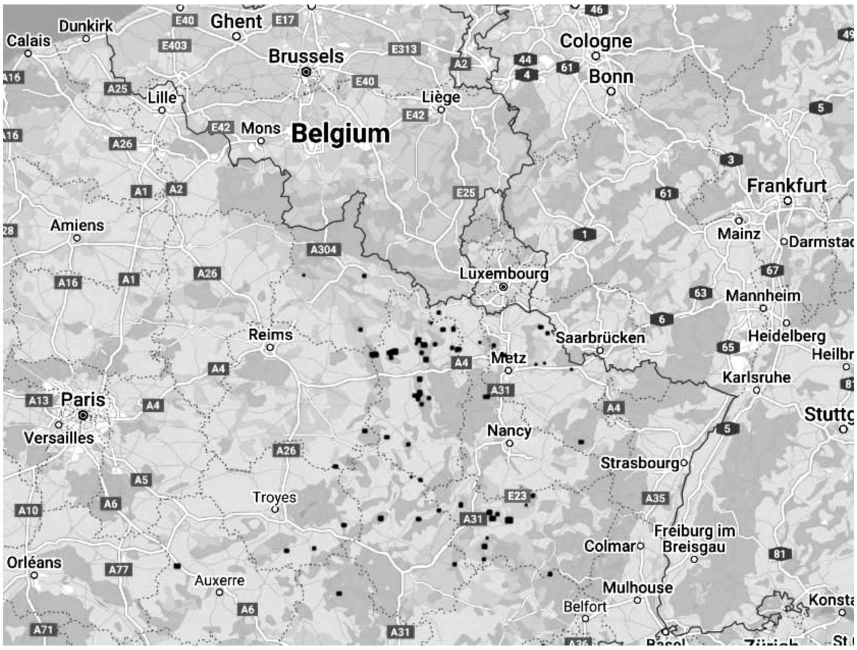


Fig. 2. Sections for which train areas are available. The locations of the areas are marked with black dots

SATELLITE DATA

The work uses multispectral images of the Sentinel-2 satellite obtained in the eastern part of France. The images contain data in 13 spectral channels (bands) with a spatial resolution of 10 to 60 meters. Images of areas of coniferous forests were selected for analysis. All channels with a resolution of 10 m (Fig. 3, *a*) and 20 m (Fig. 3, *b*) were used, as well as channel B9 (water vapor) with a resolution of 60 m (Fig. 3, *c*). This choice is due to the fact that these ranges are sensitive to the content of chlorophyll, moisture and other indicators of vegetation, the change of which may indicate the presence of diseases.

The choice of bands for research is determined by the following considerations. The B4 band (red range) is sensitive to the chlorophyll content of vegetation because chlorophyll strongly absorbs red light for photosynthesis. A decrease in the content of chlorophyll during plant stress or disease leads to an increase in reflection in the red range. B5 band (red-edge) is in the region of rapid change in reflectance from low (red light) to high (infrared). Changes in this band can carry information about the content of chlorophyll, which often changes during the development of the disease. B9 band (water vapor) is used mainly for atmospheric correction, but can help in cases of changes in the water content of vegetation under stress. Bands B11 and B12 (short-wave infrared range) are sensitive to the water content of plants because water absorbs strongly in this range. A decrease in water content under stress leads to an increase in reflectance, which may indicate the development of the disease.

In addition to the values of the spectral bands themselves, their combinations (so-called vegetation indices) will also be used as input data. Depending on the mathematical form of the index, they can highlight information about the state of the vegetation cover; eliminate or minimize the influence of negative factors (for example, brightness).

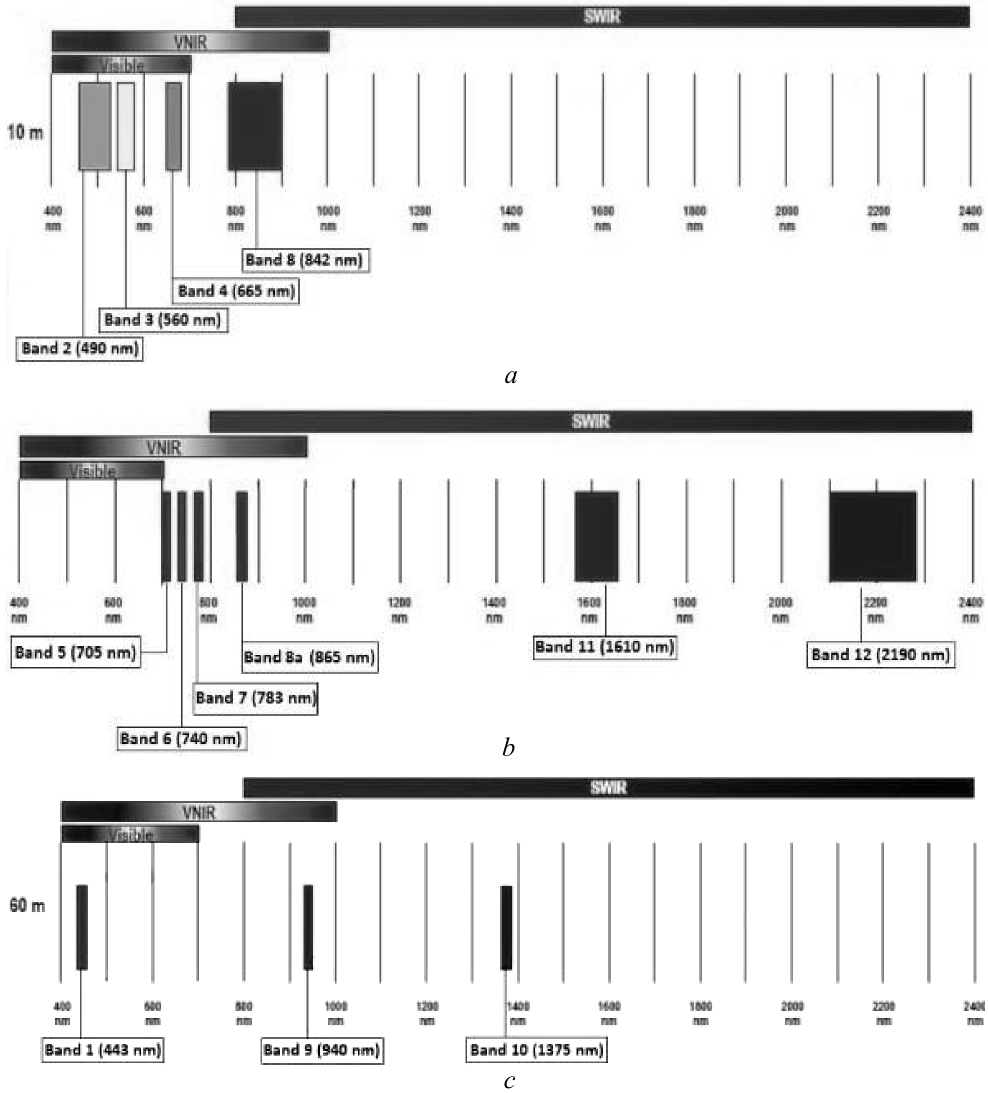


Fig. 3. Bands of Sentinel-2 satellite images [7]. Bands with a spatial resolution: 10m (*a*), 20m (*b*), 60m (*c*)

Among the well-known vegetation indices, the following can be noted:

- Green Leaf Index (GLI) — used to assess the health and development of green leaves of the plant cover, physiological state, detection of stress, drying or damage of plants, as well as monitoring of their growth and phenological changes.
- Normalized Difference Vegetation Index (NDVI) — measures the health and density of vegetation on the Earth's surface. This index is used to assess the state of ecosystems, monitor the impact of climate change and control land use.
- Disease Stress Water Index (DSWI) — used to identify plant diseases, especially coniferous forests. A decrease in DSWI values indicates a deterioration of the physiological state of the plant cover, which can be caused by diseases, for example, infectious diseases or stressful conditions.

- Chlorophyll Vegetation Index (CVI) — used to estimate the concentration of chlorophyll in the vegetation cover.

Since vegetation indices are mathematical functions, they can be generalized by classes of functions. For example, the following classes can be distinguished for the vegetation indices used in the article [2, Table 2]:

$$B(A) = A \text{ : B2, B3, B4, … , B9, B11, B12;}$$

$$NORMP(A,B) = \frac{A-B}{A+B} \text{ : NDWI, NGRDI, NDRE2, NDVI, GNDVI, NDRE3;}$$

$$FRAC(A,B) = \frac{A}{B} \text{ : RDI, PBI, CIG;}$$

$$GLIbased(A,B,C) = \frac{(A-B)+(A-C)}{(A+B)+(A+C)} \text{ : GLI;}$$

$$NORPP(A,B,C,D) = \frac{A+B}{C+D} \text{ : DSWI;}$$

$$CVIbased(A,B,C,D) = \frac{A \cdot B}{C \cdot D} \text{ : CVI;}$$

$$DIST(A,B) = \sqrt{A^2+B^2} \text{ : DRS.}$$

Set of values of various spectral bands and possible vegetation indices will be used as input data for building and training models for classifying the state of the forest into healthy and stressed.

METHODOLOGY

Bhattacharyya distance calculation

to evaluate the informativeness of the features, the bhattacharyya distance will be used between the distributions of the feature values for the healthy and damaged forest classes. The larger the value of this distance, the better the feature separates these classes.

Bhattacharyya distance is calculated by the formula:

$$D_B(H,S) = -\ln(BC(H,S)),$$

where $BC(H,S) = \sum_{i=1}^{n} \sqrt{H_i S_i}$ — Bhattacharyya coefficient, where H_i, S_i — probabilities of the *i*-th value (the height of the *i*-th columns of the histograms H and S).

The value of the Bhattacharyya coefficient is sensitive to the number of histogram columns. If their number is too low, the coefficient will be underestimated, if it is too large, it will be overestimated. Because of this, it is reasonable to take the number of histogram columns equal to the square root of the number of observations of the stressed class.

Greedy feature selection algorithm

In the previous section, classes of functions were presented, on the basis of which a large set of vegetation indices can be created. In this section, the most relevant features from this variety of indices will be defined. Our task is to select a subset of features that jointly satisfy two key criteria. First, they should be distinguished by high values of the Bhattacharyya distance that characterizes their significance. Secondly, these features should be as independent as possible, i.e. carry as much new information as possible.

This approach is aimed at building a compact and at the same time informative subset of features that helps to understand key relationships and features in the data. This selection of features will simplify the machine learning model and increase its effectiveness.

A greedy algorithm will be used to form an optimal set of informative and weakly correlated features. At each step of the work, greedy algorithms choose a locally optimal solution, which makes it possible to obtain a satisfactory global approximation in an acceptable time [8]. Heuristics based on the Bhattacharyya distance and the Spearman correlation coefficient will be used as a criterion for local optimality of the greedy algorithm.

Spearman correlation coefficient

Spearman correlation coefficient [9] is used to determine the statistical dependence of two variables and shows to what extent the dependence between variables can be described using a monotonic function. The Spearman correlation coefficient is defined as the Pearson correlation coefficient for variable ranks, that is, it does not operate with the values of quantities, but with their serial numbers. Its values lie within [–1, 1].

This value is defined as $r_s(X,Y)=\dfrac{cov(R(x),R(Y))}{\sigma(R(x))\sigma(R(y))}$, where X, Y — values of variables; $R(X)$, $R(Y)$ — transformation of variable values into their ranks; $cov(R(X),R(Y))$ — covariance of rank values; $\sigma(R(x))$, $\sigma(R(y))$ — dispersion of the corresponding rank values.

Independence coefficient

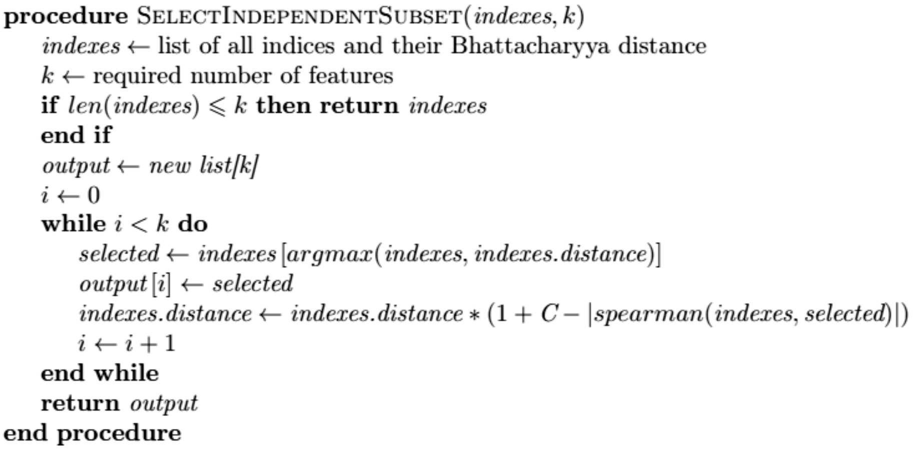
Since the goal is to find the most independent features, a value of 1 should correspond to independent features, and 0 should correspond to fully dependent features. Thus, the coefficient of independence will take the form: $c_i(X,Y)=$ $=1-|r_s(X,Y)|$, where $r_s(X,Y)$ is the Spearman correlation coefficient.

Proposed algorithm

The algorithm works as follows:

1. The weight of each feature is set equal to the value of the Bhattacharyya distance of the given feature.
2. The feature with the largest current weight is selected.
3. The weight of each feature is multiplied by the independence coefficient between the current feature and the feature selected in the previous step.
4. Return to step 2 until the required number of features have been selected

The pseudocode of the algorithm can be seen in Fig. 4. It is important to note that it uses a modified coefficient of independence, namely, a constant value C is added to it. Such change allows you to adjust what the algorithm should pay more attention to: the independence of features ($C \to 0$) or the informativeness of features ($C > 0$).



```
procedure SelectIndependentSubset($indexes, k$)
    $indexes \leftarrow$ list of all indices and their Bhattacharyya distance
    $k \leftarrow$ required number of features
    if $len(indexes) \leqslant k$ then return $indexes$
    end if
    $output \leftarrow$ new list[k]
    $i \leftarrow 0$
    while $i < k$ do
        $selected \leftarrow indexes\,[argmax(indexes, indexes.distance)]$
        $output\,[i] \leftarrow selected$
        $indexes.distance \leftarrow indexes.distance * (1 + C - |spearman(indexes, selected)|)$
        $i \leftarrow i + 1$
    end while
    return $output$
end procedure
```

Fig. 4. Pseudocode of the proposed algorithm

This approach allows to consistently add to the set the most informative at the moment and weakly correlated with previous signs. As a result, a set is formed that contains a maximum of information for a given number of features.

Machine learning models for classification

To compare the effectiveness of different sets of features, machine learning models will be used for binary classification of the forest state.

Multilayer perceptron [10] with a different number of hidden layers will be used as a basic classifier. Optimal architecture and hyperparameters of the model will be tuned using a genetic algorithm.

The models will be trained on a training sample of satellite images with ground truth masks. To assess the quality of the classification, such metrics as [11] will be used: overall accuracy (accuracy), Jacquard coefficient (IoU), cross-entropy (log-loss), area under the ROC curve (ROC AUC) on the validation sample.

Note that the overall accuracy metric makes sense only when assessing the classification accuracy with a balanced distribution of classes, which occurred on the validation sample.

Five-fold cross-validation was used to analyze metrics.

Comparing the results of models built on different sets of features allows us to assess the contribution of the proposed feature selection method to improving the quality of classification.

EXPERIMENT RESULTS

Description and results of the experiment on evaluating the informativeness of features

In order to evaluate the informativeness of various features of the image, the Bhattacharyya distance between the classes of “stressed” and “healthy” coniferous forest was calculated for various spectral channels and vegetation indices.

Sentinel-2 images containing fragments of healthy and stressed coniferous forest in eastern France were used as input data. Each image contains 12 bands.

For each pixel of each image, 43.644 vegetation indices were calculated based on 12 spectral channels according to the given index classes.

The territories were divided into 3 parts, for each of which Bhattacharyya distances were calculated separately between the obtained histograms of class distributions. Histograms were built within the values of the "stressed" class with the number of columns equal to the square root of the number of pixels of this class.

As a result, average estimates of the Bhagattacharya distance ($Avg(D_B(H,S))$) for each of the features were obtained.

Among the original spectral channels, bands B4, B11 and B12 showed the greatest informativeness (Table 1). Among the known vegetation indices, 7 (RDI, NDWI, NGRDI, DSWI, NDRE2, NDVI, GLI) demonstrated higher separation rates (Table 1) compared to spectral channels. A comparison of Bhattacharyya distance values in vegetation indices from table 1 and the corresponding class of indices (Table 2) shows that there are instances of the class with a larger distance value.

Table 1. Bhattacharyya distance

Sentinel-2 bands		**Well-known vegetation indices**		
Band	$Avg(D_B(H,S))$	Index	Formula	$Avg(D_B(H,S))$
$B12$	0.332	RDI	$\frac{B12}{B8A}$	0.581
$B4$	0.306	NDWI	$\frac{B8A-B11}{B8A+B11}$	0.577
$B11$	0.228	NGRDI	$\frac{B3-B4}{B3+B4}$	0.562
$B1$	0.116	DSWI	$\frac{B8+B3}{B4+B11}$	0.513
$B9$	0.115	NDRE2	$\frac{B7-B5}{B7+B5}$	0.470
$B5$	0.107	NDVI	$\frac{B8A-B4}{B8A+B4}$	0.448
$B7$	0.092	GLI	$\frac{(B3-B4)+(B3-B2)}{(B3+B4)+(B3+B2)}$	0.389
$B2$	0.073	PDI	$\frac{B8}{B3}$	0.219
B8A	0.068	CIG	$\frac{B8A}{B3}-1$	0.183
$B6$	0.068	GNDVI	$\frac{B8A-B3}{B8A+B3}$	0.182
$B8$	0.062	NDRE3	$\frac{B8A-B7}{B8A+B7}$	0.126
$B3$	0.041	CVI	$\frac{B8A\cdot B5}{B3^2}$	0.051

Table 2. The largest Bhattacharyya distance for classes of vegetation indices

Class	**Instance**	$Avg(D_B(H,S))$
$CVIbased(A,B,C,D)$	$\frac{B3 \cdot B6}{B11 \cdot B4}$	0.686
$NORPP(A,B,C,D)$	$\frac{B2+B6}{B12+B4}$	0.646
$GLIbased(A,B,C)$	$\frac{(B6-B4)+(B6-B11)}{(B6+B4)+(B6+B11)}$	0.630
$NORMP(A,B)$	$\frac{B6-B12}{B6+B12}$	0.609
$FRAC(A,B)$	$\frac{B6}{B12}$	0.603
$DIST(A,B)$	$\sqrt{B12^2+B4^2}$	0.354
$B(A)$	$B12$	0.331

So, the experiment confirmed that the Bhattacharyya distance can be used to evaluate the informativeness of features, confirmed the advantage of using vegetation indices in comparison with the original image data, and showed that for each class of indices it is possible to find instances with better informativeness (within the scope of the task) than in well known indices.

This makes it possible to form an effective set of features for further training of machine learning models without the need for their previous training.

Determination of the optimal set of informative features

The analysis of the results of feature selection showed that among the initial set of 43.644 calculated vegetation indices, 3.128 had a Bhattacharyya distance above 0.4, which indicates their high informativeness.

Using the proposed greedy algorithm, sets of 12 and 24 features were obtained. The algorithm used a modified coefficient of independence: $c_i(X,Y)+C$, where $C=0.6$. At this value of the C parameter, the model showed the best result.

As it was found during experiments, the use of classes $GLIbased(A,B,C)$, $NORPP(A,B,C,D)$, $CVIbased(A,B,C,D)$ during selection leads to a decrease in metric values, so their use was abandoned. Rejecting them allows you to reduce the execution time of the algorithm by an order of magnitude.

As a result of this problem, it seems reasonable to search for an effective combination within each class separately, and then somehow combine them. However, the identification of the causes of this problem and methods of solvingit require a separate study.

As can be seen in Fig. 5, the obtained set of features mostly contains features with relatively little informativeness. This indicates that although a large number of signs are informative, they are also highly correlated. This is also confirmed by the fact that the selected features with high informativeness are more correlated with each other (have a darker color) than the features with low.

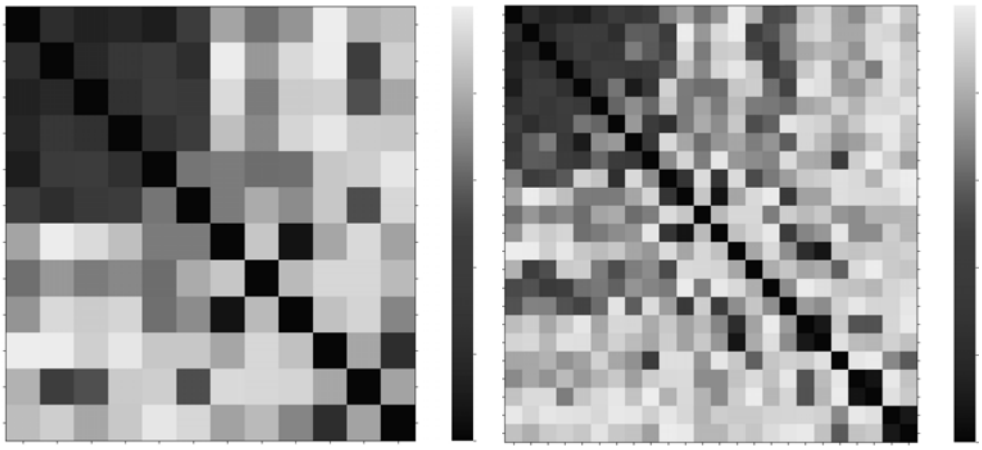


Fig. 5. Independence matrix for (*a*) — 12, (*b*) — 24 features selected by the proposed algorithm among the classes *NORMP*(*A*, *B*), *FRAC*(*A*, *B*), DIST(*A*, *B*)

Fig. 5 shows the example of pairwise independence of selected features, where features are placed from top-left to bottom-right in descending order of their individual informativeness. Brightness of each cell corresponds to their independence, where brighter means higher.

Overall, the obtained set of features mostly contains features with relatively little informativeness. As can be seen in fig. 5, the selected features with high informativeness are more correlated with each other (have a darker color) than the features with low.

Analysis of machine learning results

To evaluate how successful the sets turned out to be, let's compare the accuracy of models built using them compared to models using only spectral channels, known vegetation indices, and their combination.

Fig. 6 immediately shows that the model built on the features proposed by the algorithm shows much better knowledge of metrics and their dynamics. Thus, already after about 7 epochs, the model based on 12 proposed features shows metrics close to the metrics of the model based on known vegetation indices. This comparison confirms that the proposed algorithm is effective.

It should also be noted that the models from the 12 proposed features show close (but still slightly worse) values of metrics to the model based on spectral classes. At the same time, the algorithm worked an order of magnitude longer than model training. Based on this, it can be concluded that the spectral channels are a good enough basis, and the multilayer perceptron model is able to build a vegetative index with high informativeness on their basis.

However, when the number of features is doubled, the proposed algorithm was able to find such a set of features that improves the accuracy of the model and the rate of its learning. This suggests that the feature sets previously failed to maximize the amount of information, which one would think, as the model based on a combination of known vegetation indices and spectral channels showed almost the same accuracy and learning rate as the model based on spectral channels alone. And also confirms the usefulness and necessity of feature engineering.

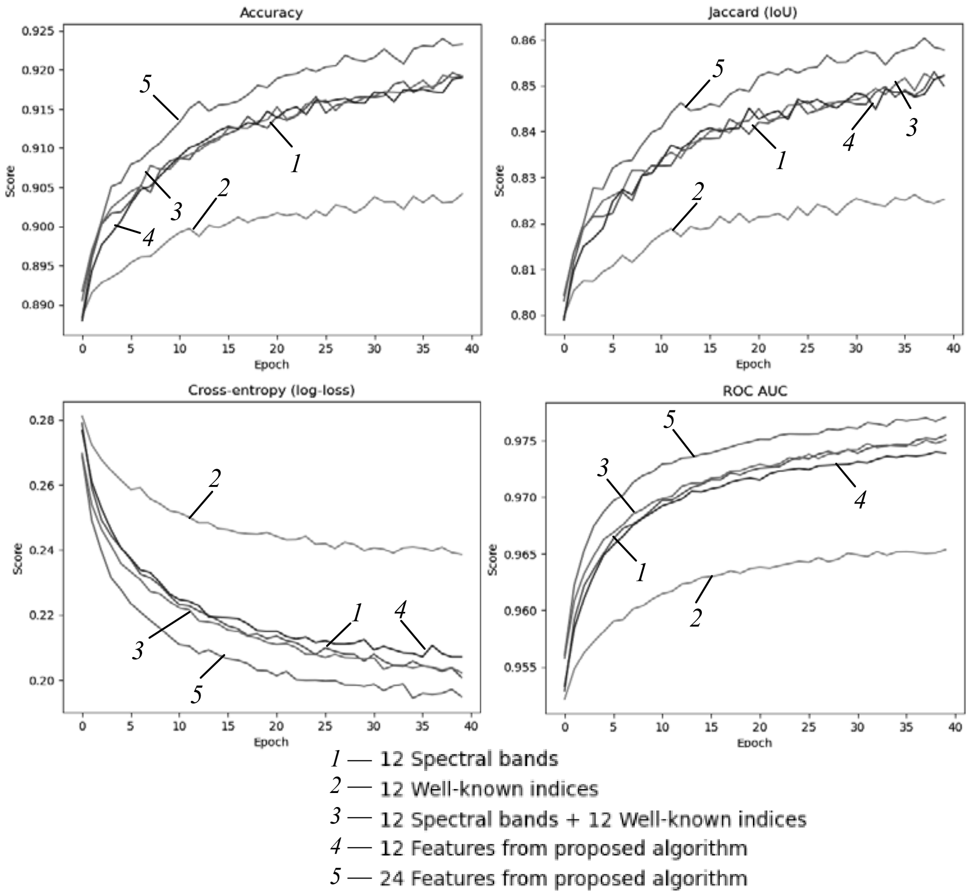


Fig. 6. Metrics of built models on the validation sample

In the future, the proposed algorithm can be applied to optimize forest state classification models based on other types of data and for other tasks of remote sensing of the Earth.

CONCLUSIONS

In this work, the possibility of using Bhattacharyya distance to assess the relative importance of features in the task of forest state classification based on satellite images was investigated.

The analysis of real data showed that it makes sense to consider not specific (known) vegetation indices, but their classes. It was confirmed that within each class, with a high probability, a vegetative index can be found that is more informative compared to the known and original Sentinel-2 spectral channels.

The proposed greedy feature selection algorithm based on the Bhattacharyya distance and the Spearman correlation coefficient made it possible to form a set of 12 features with similar accuracy indicators, and a set of 24 features with significantly better ones, compared to the model based only on the spectral channels of the image.

Therefore, the proposed approach is effective for selecting informative and weakly correlated features based on satellite images. It can be applied to find an effective set of features for building machine learning models in forest condition monitoring tasks and other fields of Earth remote sensing data analysis without the need to pre-train the models.

REFERENCES

1. N. Kussul, G. Lemoine, J. Gallego, S. Skakun, and M. Lavreniuk, "Parcel based classification for agricultural mapping and monitoring using multi-temporal satellite image sequences," *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2015*. doi: 10.1109/igarss.2015.7325725.
2. J. Zhang, S. Cong, G. Zhang, Y. Ma, Y. Zhang, and J. Huang, "Detecting Pest-Infested Forest Damage through Multispectral Satellite Imagery and Improved UNet++," *Sensors*, vol. 22, issue 19, 2022. doi: 10.3390/s22197440.
3. N.N. Kussul, N.S. Lavreniuk, A.Y. Shelestov, B.Y. Yailymov, and I.N. Butko, "Land Cover Changes Analysis Based on Deep Machine Learning Technique," *Journal of Automation and Information Sciences*, vol. 48, no. 5, pp. 42–54, 2016. doi: 10.1615/jautomatinfscien.v48.i5.40.
4. T. van Erven, P. Harrëmos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, 60(7), 2014. Available: https://doi.org/10.1109/TIT.2014.2320500
5. A. Ilnitskiy, O. Burba, "Statistical criteria for assessing the informativity of the sources of radio emission of telecommunication networks and systems in their recognition," *Cybersecurity: Education, Science, Technique*, 1(5), pp. 83–94, 2019. doi: 10.28925/2663-4023.2019.5.8394.
6. *Forest type 2018*. Accessed on: April 07, 2023. [Online]. Available: https://land.copernicus. eu/pan-european/highresolution-layers/forests/forest-type-1/status-maps/forest-type-2018.
7. "Spatial Resolutions," *Sentinel Online*. Accessed on: August 13, 2023. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi /resolutions/spatial
8. "What is a Greedy Approach? - Algorithms for Coding Interviews in Java," *educative.io*. Accessed on: May 08, 2023. [Online]. Available: https://www.educative.io/courses/algorithms-coding-interviews-java/3j1R50KnNjQ
9. C. Croux, C. Dehon, "Influence functions of the Spearman and Kendall correlation measures," *Statistical Methods and Applications*, vol. 19, pp. 497–515, 2010. doi: 10.1007/s10260-010-0142-z.
10. P.M. Atkinson, A.R. Tatnall, "Introduction neural networks in remote sensing," *International Journal of Remote Sensing*, vol. 18(4), 1997. doi: 10.1080/014311697218700.
11. "Metrics for semantic segmentation," *ilmonteux.github.io*. Accessed on: May 27, 2023. [Online]. Available: https://ilmonteux.github.io/2019/05/10/segmentation-metrics.html

Received 06.09.2023

INFORMATION ON THE ARTICLE

Yevhenii V. Salii, ORCID: 0009-0006-0395-8099, Educational and Research Institute of Physics and Technology of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: yevhenii.salii@gmail.com

Alla M. Lavreniuk, ORCID: 0000-0002-5791-0377, Educational and Research Institute of Physics and Technology of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: alla.lavrenyuk@gmail.com

Nataliia M. Kussul, ORCID: 0000-0002-9704-9702, Educational and Research Institute of Physics and Technology of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: nataliia.kussul@gmail.com

СТАТИСТИЧНІ МЕТОДИ ІНЖЕНЕРІЇ ОЗНАК ДЛЯ ЗАДАЧІ КЛАСИФІКАЦІЇ СТАНУ ЛІСІВ ЗА СУПУТНИКОВИМИ ДАНИМИ / Є.В. Салій, А.М. Лавренюк, Н.М. Куссуль

Анотація. Своєчасне виявлення хвороб лісу є важливим завданням для запобігання їх поширенню та обмеження наслідків. Використання супутникових зображень надає можливості для великомасштабного моніторингу лісів. Моделі машинного навчання дають змогу автоматизувати аналіз цих даних для виявлення аномалій, що можуть свідчити про наявність хвороб. Відбір інформативних ознак є ключовим етапом побудови ефективної моделі. Досліджено можливість застосування відстані Бгаттачар’я та коефіцієнта кореляції Спірмена для відбору ознак із супутникових зображень. Застосовано жадібний алгоритм для формування підмножини слабко корельованих ознак. Експеримент показав, що обрані ознаки дозволяють покращити якість класифікації порівняно з використанням усіх спектральних каналів. Запропонований підхід продемонстрував ефективність для відбору інформативних і слабко корельованих ознак та може застосовуватися в інших задачах дистанційного зондування Землі.

Ключові слова: Sentinel-2, вегетаційні індекси, відстань Бгаттачар’я, інженерія ознак, жадібні алгоритми, коефіцієнт кореляції Спірмена.