# COMPARING DIFFERENT MACHINE LEARNING OPTIONS TO MAP BARK BEETLE INFESTATIONS IN CROATIA

N. Kranjčić[1*], V. Cetl[2], H. Matijević[2], D. Markovinović[2]

[1] Faculty of Geotechnical Engineering, University of Zagreb, Varaždin, Croatia - nikola.kranjcic@gfv.unizg.hr
[2] Department for Geodesy and Geomatics, University North, Varaždin, Croatia - (vcetl, hmatijevic, dmarkovinovic)@unin.hr

**KEY WORDS:** supervised classification, machine learning options, QGIS, SAGA GIS, Copernicus data

**ABSTRACT:**

This paper presents different approaches to map bark beetle infested forests in Croatia. Bark beetle infestation presents threat to forest ecosystems. Due to large unapproachable area, it also presents difficulties in mapping infested areas. This paper analyses available machine learning options in open-source software QGIS and SAGA GIS. All options are performed on Copernicus data, Sentinel 2 satellite imagery. Machine learning and classification options are maximum likelihood classifier, minimum distance, artificial neural network, decision tree, K Nearest Neighbor, random forest, support vector machine, spectral angle mapper and Normal Bayes. Kappa values respectively are: 0.71; 0.72; 0.81; 0.68; 0.69; 0.75; 0.26; 0.60; 0.41 which shows highest classification accuracy for artificial neural networks method and lowest for support vector machine accuracy.

## 1. INTRODUCTION

Remote sensing is the process of acquiring information about an object or phenomenon without making physical contact with it. It involves usage of various sensors to capture data from a distance, such as aerial photography, satellite imagery, and LiDAR. One of the most important applications of remote sensing is classification. It is the process of categorizing objects or areas based on their characteristics in the acquired data. In recent years, machine learning methods have become increasingly popular for remote sensing classification. Machine learning algorithms, such as artificial neural networks, support vector machines, and random forests, are used to automatically learn and recognize patterns in the data, and then assign classification labels to the objects or areas. These methods have been shown to be effective for a wide range of remote sensing applications, including land use and land cover mapping, vegetation monitoring, and urban growth analysis. Within this paper we explore the use of machine learning methods for remote sensing data classification (Feng et al., 2015; Foody, 2002; Jain et al., 2016; Jog and Dixit, 2016; Kranjčić et al., 2019a; Singh et al., 2017). We discuss various algorithms, their strengths and weaknesses, and their suitability for different types of remote sensing data. We also investigate the impact of different input features, such as spectral, textural, and contextual information, on the performance of the classifiers. Finally, we compare the results of different machine learning methods with traditional classification techniques and discuss the potential for future research and development in this field. However, due to the page limitations, each method and comparations are defined partially. Following methods are used and discussed: maximum likelihood, minimum distance, artificial neural network, decision tree, K nearest neighbour, random forest, support vector machine, spectral angle mapper and naïve Bayes. We executed all the classification methods using OpenCV library from within QGIS and SAGA GIS software. The paper is organised as it follows, second chapter presents methods, study area and data sets used. Chapter three deals with results and discussion, chapter four presents' conclusions and lest chapter shows references used.

## 2. METHODS, STUDY AREA AND DATA SETS

In this chapter, each method is shortly explained, study area is presented together with data sets used.

### 2.1 Maximum likelihood

Maximum likelihood (ML) is a supervised classification method based on Bayes theorem. It is a statistical method that uses probability theory to classify each pixel in an image into different land cover categories based on its spectral characteristics (Ahmad and Quegan, 2012).

### 2.2 Minimum distance

Minimum distance (MD) classifier depends on training data used to perform classification on unknown data set to the classes that minimizes distance between images and classes in multidimensional space. Minimum distance shows maximum similarity. Due to smaller number of calculations it requires less processing time (Jog and Dixit, 2016).

### 2.3 Artificial neural network

Artificial neural networks (ANNs) are a type of machine learning algorithm inspired by the structure and function of biological neurons. ANNs have been widely used in remote sensing classification due to their ability to learn complex relationships between input variables and output classes, and their ability to handle non-linear relationships in the data. ANNs are composed of multiple layers of interconnected nodes, or neurons, which receive input signals, perform a non-linear transformation on those signals, and then pass the transformed signals to the next layer of neurons. The final layer of neurons produces the output, which is typically a classification label (Miller et al., 1995; Song et al., 2012).

### 2.4 Decision tree

Decision tree (DT) is a general, predictive modelling tool with applications in different areas. Decision trees are constructed via an algorithmic approach that describes ways to split a data set based on specific tasks. Due to method simplicity, it is one

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-83-2023

of the most widely used and practical methods for supervised learning. It is a non-parametric supervised learning method used for both classification and regression tasks (Kumar, 2022; Song and Lu, 2015).

### 2.5 K-nearest neighbor

K-nearest neighbor (KNN) is one of the most basic yet significant classification algorithms in machine learning. It is a supervised machine learning method often used in the domain of pattern recognition, data mining and intrusion detection. It can solve classification and regression problems (Meng et al., 2007).

### 2.6 Random forest

Random forest (RF) is a popular machine learning algorithm used for remote sensing classification that combines multiple decision trees to improve classification accuracy and reduce overfitting. In the RF algorithm, multiple decision trees are trained on different subsets of the training data and with a random subset of input features. Each tree makes a classification decision based on the selected features, and the final classification decision is made by aggregating the decisions of all the trees through a majority voting scheme (Kranjčić et al., 2019a; Oliveira et al., 2012; Pal, 2005; Rodriguez-Galiano et al., 2015).

### 2.7 Support vector machine

Support vector machine (SVM) is a supervised learning algorithm that seeks to find a hyperplane that separates the data into different classes with the largest margin between the classes. In SVM, each pixel in the image is represented as a point in a high-dimensional space, and the algorithm seeks to find the hyperplane that best separates the different classes. The hyperplane is selected to maximize the margin between the closest points of the different classes, which are known as support vectors (Jog and Dixit, 2016; Kranjčić et al., 2019b; Naghibi et al., 2017; Ngoc Thach et al., 2018).

### 2.8 Spectral angle mapper

Spectral angle mapper (SAM) calculates the spectral angle between a pixel's spectral signature and the spectral signature of a known target class to determine its class membership. SAM assumes that the spectral signatures of different materials can be represented as vectors in a high-dimensional space, and that the angle between two vectors represents the similarity between the two spectra. SAM calculates the angle between the pixel's spectral signature and the spectral signature of each target class and assigns the pixel to the class with the smallest angle (De Carvalho and Meneses, 2000; Liu and Yang, 2013).

### 2.9 Naïve Bayes

Naive Bayes (NB) is a probabilistic machine learning algorithm based on Bayes' theorem, which describes the probability of a hypothesis given some observed evidence. In Naive Bayes, each pixel in the image is represented as a vector of input features, such as spectral bands or texture measures. The algorithm assumes that each input feature is independent of the others, which is known as the "naive" assumption. Naive Bayes calculates the probability of each class given the input features and assigns the pixel to the class with the highest probability. The probability of each class is calculated using Bayes' theorem, which incorporates the prior probability of the class and the probability of the input features given the class. (Kholod et al., 2019; Solares and Sanz, 2005; Soria et al., 2011; Wieland and Pittore, 2014)

### 2.10 Study area

Study area is in the in mountainous area of Croatia, where spruce, beech and fir trees can be found. Municipality of Čabar is located on altitude 650 to 1200 meters above sea level and it is covered with spruce and fir forests. During 2014 bark beetle infestation outbreak was registered at municipality of Čabar. Spruce forests are infected with bark beetles and main characteristics are yellow/red treetops which can be distinguished on remote sensed data (Kranjčić et al., 2018).

### 2.11 Used data sets

We used Copernicus Sentinel 2A multispectral images. Sentinel 2A contains multispectral imager covering 13 spectral bands (443nm – 2190 nm) with spatial resolution of 10 m, 20 m and 60m (Agency, 2021; Rättich et al., 2020). Date of downloaded data is 04th August 2017. Figure 1 shows study area, training, and control data sets.

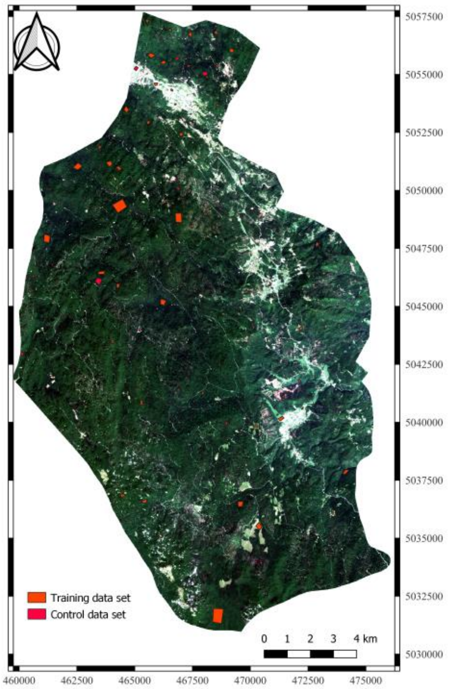

**Figure 1**. Study area, training, and control data.

### 2.12 Accuracy assessment

Viera et al. (2005) shows that kappa analysis is a powerful tool to compare differences between classification results. Kappa values between 0.41 and 0.60 indicate that classification is moderate accuracy. Kappa values between 0.61 and 0.80 shows high accuracy and kappa values higher than 0.80 indicates very high classification accuracy.

## 3. RESULTS AND DISCUSSION

Table 1 shows kappa values for each method. Figures 2 to 10 show results of supervised classification for each above-mentioned method, as it follows: maximum likelihood, minimum distance, artificial neural network, decision tree, K-nearest neighbor, random forest, support vector machine, spectral angle mapper and naïve Bayes.

| Method | Kappa value |
|---|---|
| Maximum likelihood | 0.71 |
| Minimum distance | 0.72 |
| Artificial neural network | 0.81 |
| Decision tree | 0.68 |
| K-nearest neighbor | 0.69 |
| Random forest | 0.75 |
| Support vector machine | 0.26 |
| Spectral angle mapper | 0.60 |
| Naïve Bayes | 0.41 |

**Table 1**. Kappa values.

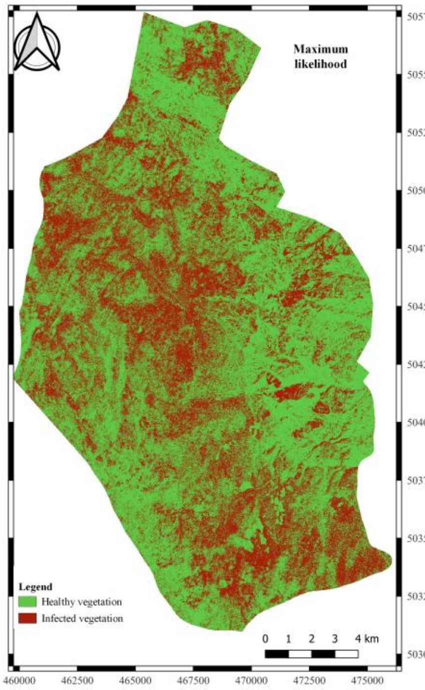

**Figure 2**. Results of maximum likelihood classification.

As indicated in Table 1 and Figures 2-10 support vector machine classification is low accuracy, and deviates from other methods. Other methods have similar kappa values, therefore similar classification accuracy. Highest results are achieved using artificial neural networks with kappa value 0.81. Training data set, control data set, number of data sets and size of specific sample have effect on classification results. However, for many methods it seems that all parameters have similar effects. This shows that further research needs to be done to establish connections between parameters and classification results.

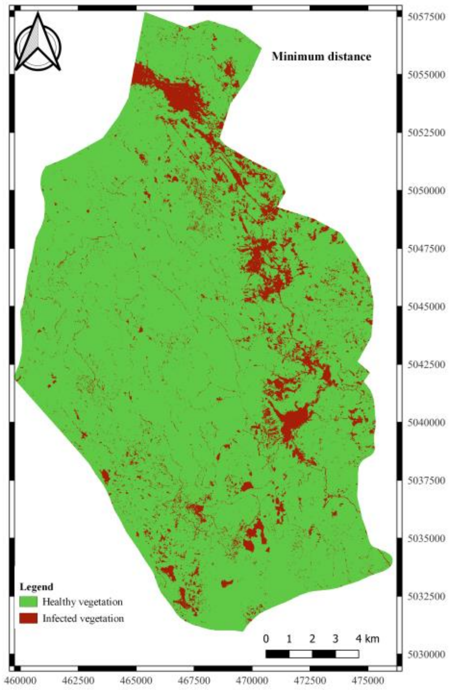

**Figure 3**. Results of minimum distance classification.

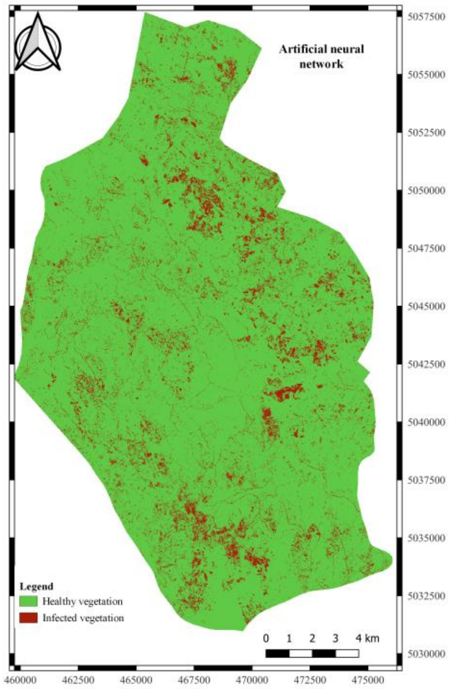

**Figure 4**. Results of artificial neural network classification.

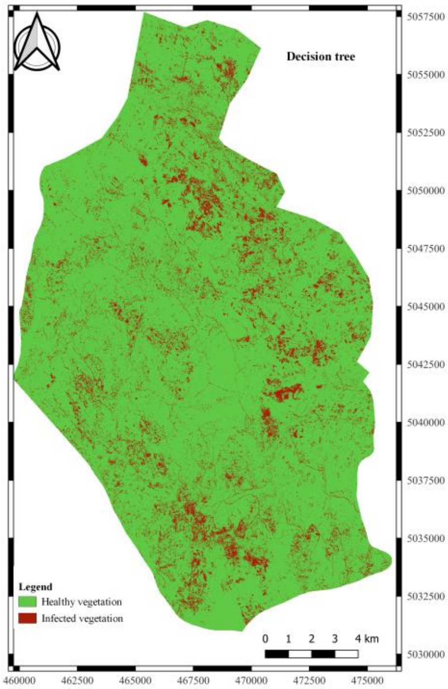

**Figure 5**. Results of decision tree classification.

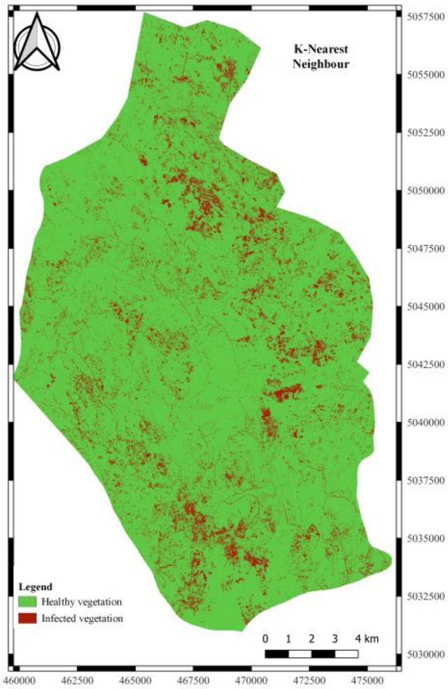

**Figure 6**. Results of K-nearest neighbour classification.

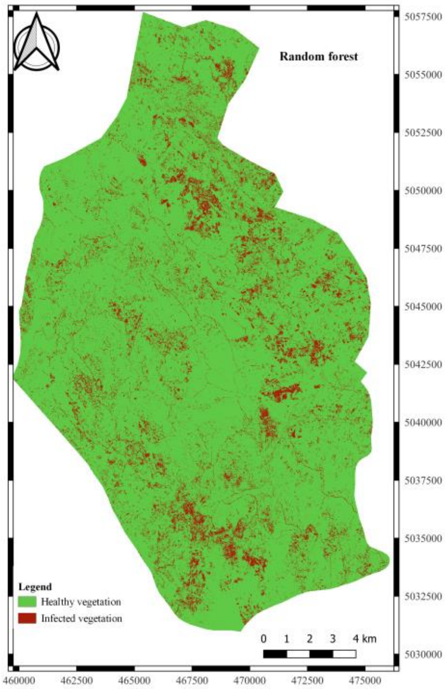

**Figure 7**. Results of random forest classification.

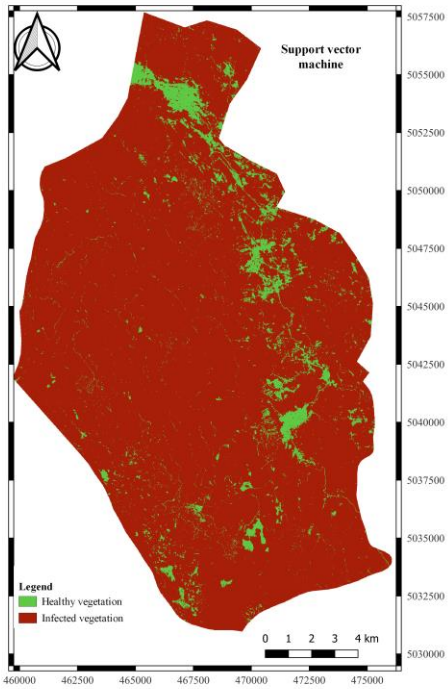

**Figure 8**. Results of support vector machine classification.

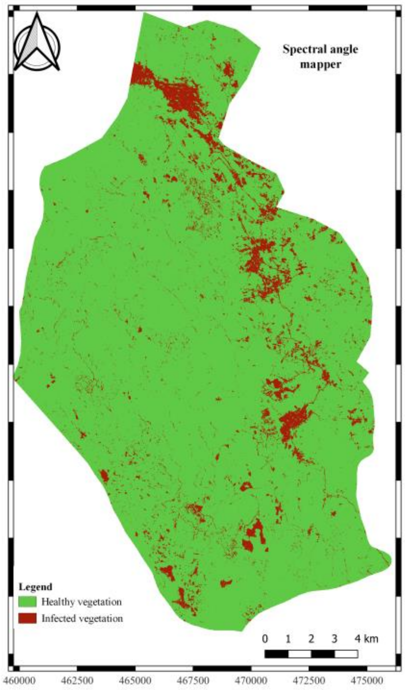

**Figure 9**. Results of spectral angle mapper classification.

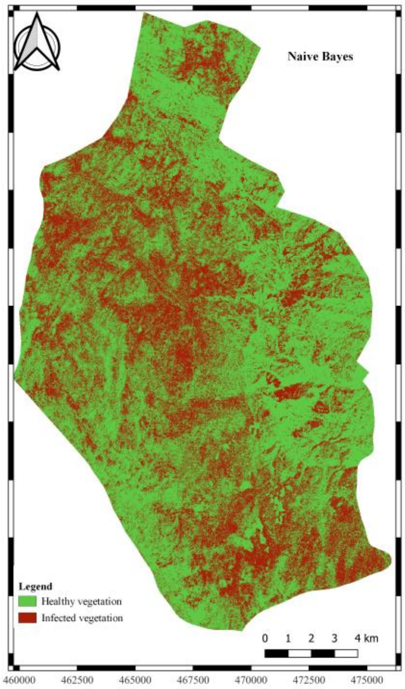

**Figure 10**. Results of naïve Bayes classification.

## 4. CONCLUSIONS

In this paper we used several machine learning methods for bark beetle infestation mapping. For classification we used QGIS and SAGA GIS software and all methods are based on OpenCV library. Methods analyzed are as follows: maximum likelihood, minimum distance, artificial neural network, decision tree, K-nearest neighbor, random forest, support vector machine, spectral angle mapping and naïve Bayes. Kappa values respectively are: 0.71; 0.72; 0.81; 0.68; 0.69; 0.75; 0.26; 0.60; 0.41. This indicates that artificial neural networks achieved highest classification accuracy and support vector machine accuracy is the lowest. Such results were expected, however higher classification accuracy for support vector machine should be achieved. Results are influenced by various parameters such as training data set, control data set, number of data sets and size of specific sample. Therefore, future research must include exploration how specific parameter affects classification accuracy.

## ACKNOWLEDGEMENTS

This research is supported by the scientific project Digital twins and smart cities from the University North, Croatia.

## REFERENCES

European Space Agency, 2021. Sentinel-2 Eur. Sp. Agency. sentinels.copernicus.eu/web/sentinel/missions/sentinel-2 (24 May 2023).

Ahmad, A., Quegan, S., 2012. Analysis of maximum likelihood classification on multispectral data. *Appl. Math. Sci.* 6, 6425–6436.

De Carvalho, O.A., Meneses, P.R., 2000. Spectral correlation mapper (SCM): an improvement on the spectral angle mapper (SAM), in: *Summaries of the 9th JPL Airborne Earth Science Workshop, JPL Publication* 00-18. p. 2.

Feng, Q., Liu, J., Gong, J., 2015. UAV Remote Sensing for Urban Vegetation Mapping Using Random Forest and Texture Analysis. *Remote Sens.* . doi.org/10.3390/rs70101074.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sens. Environ.* 80, 185–201. doi.org/10.1016/S0034-4257(01)00295-4.

Jain, M., Dawa, D., Mehta, R., Dimri, A.P., Pandit, M.K., 2016. Monitoring land use change and its drivers in Delhi, India using multi-temporal satellite data. *Model. Earth Syst. Environ.* 2, 19. doi.org/10.1007/s40808-016-0075-0.

Jog, S., Dixit, M., 2016. Supervised classification of satellite images. *Conf. Adv. Signal Process. CASP* 2016 93–98. doi.org/10.1109/CASP.2016.7746144.

Kholod, I.I., Kuprianov, M.S., Titkov, E. V, Shorov, A. V, Postnikov, E. V, Mironenko, I.G., Sokolov, S.S., 2019. Training Normal Bayes Classifier on Distributed Data. *Procedia Comput. Sci.* 150, 389–396. doi.org/10.1016/j.procs.2019.02.068.

Kranjčić, N., Medak, D., Župan, R., Rezo, M., 2019a. Machine Learning Methods for Classification of the Green Infrastructure in City Areas. *ISPRS Int. J. Geo-Information* . doi.org/10.3390/ijgi8100463.

Kranjčić, N., Medak, D., Župan, R., Rezo, M., 2019b. Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns. *Remote Sens.* . doi.org/10.3390/rs11060655.

Kranjčić, N., Župan, R., Rezo, M., 2018. Satellite-based hyperspectral imaging and cartographic visualization of bark beetle forest damage for the city of Čabar. *Teh. Glas.* 12, 39–43. doi.org/10.31803/tg-20171219085721.

Kumar, V., 2022. Decision Tree Algorithm overview explained towardsmachinelearning. URL https://towardsmachinelearning.org/decision-tree-algorithm/ (16 May 2023).

Liu, X., Yang, C., 2013. A Kernel Spectral Angle Mapper algorithm for remote sensing image classification. *Proc. 2013 6th Int. Congr. Image Signal Process.* CISP 2013 2, 814–818. doi.org/10.1109/CISP.2013.6745277.

Meng, Q., Cieszewski, C.J., Madden, M., Borders, B.E., 2007. K nearest neighbor method for forest inventory using remote sensing data. *GIScience Remote Sens.* 44, 149–165. doi.org/10.2747/1548-1603.44.2.149.

Miller, D.M., Kaminsky, E.J., Rana, S., 1995. Neural network classification of remote-sensing data. *Comput. Geosci*. 21, 377–386. doi.org/10.1016/0098-3004(94)00082-6.

Naghibi, S.A., Ahmadi, K., Daneshi, A., 2017. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in *Groundwater Potential Mapping. Water Resour. Manag*. 31, 2761–2775. doi.org/10.1007/s11269-017-1660-3.

Ngoc Thach, N., Bao-Toan Ngo, D., Xuan-Canh, P., Hong-Thi, N., Hang Thi, B., Nhat-Duc, H., Dieu, T.B., 2018. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecol. Inform*. 46, 74–85. doi.org/https://doi.org/10.1016/j.ecoinf.2018.05.009.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J.M.C., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manage*. 275, 117–129. doi.org/https://doi.org/10.1016/j.foreco.2012.03.003.

Pal, M., 2005. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* 26, 217–222. doi.org/10.1080/01431160412331269698.

Rättich, M., Martinis, S., Wieland, M., 2020. Automatic Flood Duration Estimation Based on Multi-Sensor Satellite Data. *Remote Sens.* . doi.org/10.3390/rs12040643.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* 71, 804–818. doi.org/https://doi.org/10.1016/j.oregeorev.2015.01.001.

Singh, S.K., Kumar, V., Kanga, S., 2017. Land Use/Land Cover Change Dynamics and River Water Quality Assessment Using Geospatial Technique: a case study of Harmu River, Ranchi (India) for the Sustainable Management of Water Resources View project Land Use/Land Cover Change Dynamics and River Wat. *Int. J. Sci. Res. Comput. Sci. Eng*. 5, 17–24.

Solares, C., Sanz, A.M., 2005. Bayesian network classifiers. An application to remote sensing image classification. *WSEAS Trans. Syst.* 4, 343–348.

Song, K.-Y., Oh, H.-J., Choi, J., Park, I., Lee, C., Lee, S., 2012. Prediction of landslides using ASTER imagery and data mining models. *Adv. Sp. Res*. 49, 978–993. doi.org/https://doi.org/10.1016/j.asr.2011.11.035.

Song, Y.Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* 27, 130–135. doi.org/10.11919/j.issn.1002-0829.215044.

Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O., 2011. A ‘non-parametric’ version of the naive Bayes classifier. *Knowledge-Based Syst.* 24, 775–784. doi.org/https://doi.org/10.1016/j.knosys.2011.02.014.

Viera, A.J., Garrett, J.M., others, 2005. Understanding interobserver agreement: the kappa statistic. *Fam med 37*, 360–363.

Wieland, M., Pittore, M., 2014. Performance Evaluation of Machine Learning Algorithms for Urban Pattern Recognition from Multi-spectral Satellite Images. *Remote Sens.* . doi.org/10.3390/rs6042912.