Creating Models of Custom Image Classification Workflows Using Machine Learning Techniques

JASON D. WOLFE[1], THOMAS BAHR[2] & NICOLAI HOLZER[2]

Abstract: Machine learning algorithms are becoming increasingly popular for image classification due to their robustness and accuracy. This study demonstrates a streamlined process for comparing the accuracy of different machine learning classifiers using a visual modelling tool. The classifiers included ensemble learning methods such as AdaBoost, Random Forest, and Extra Trees; along with traditional methods such as Maximum Likelihood and Minimum Distance. Classification models were built and run using Sentinel-2A imagery of a study area with vegetation land-cover classes that were spectrally similar. The Random Forest and Extra Trees classifiers yielded overall accuracies of 100% for 145,514 training samples. The AdaBoost classifier yielded an overall accuracy of 93% for 210,235 samples. The traditional Maximum Likelihood and Minimum Distance yielded lower accuracies: 84% and 45% for 147,601 samples.

1 Introduction

With the continued advancement of new machine learning algorithms, the options that are available for image classification have increased significantly. Comparing the performance and accuracy of multiple classification algorithms can be time-consuming, particularly if analysts choose to write programming (API) code. This study shows how a visual modelling tool can provide a streamlined process to building custom image classification workflows.

Visual models allow for easy customization by interchanging different trainers and classifiers. Inputs, outputs, data management operations, and processing tasks can be linked with a drag-and-drop user interface instead of learning API code.

A case study demonstrates the use of simple models to run multiple supervised classification workflows with Sentinel-2A multispectral imagery. The workflows use three machine learning algorithms from the Python "scikit-learn" package (PEDREGOSA et al. 2011):

- AdaBoost
- Extra Trees
- Random Forest

Two additional workflows were built and run using Minimum Distance and Maximum Likelihood classifiers, which are traditional remote sensing algorithms.

[1] Harris Geospatial Solutions Inc, 385 Interlocken Crescent Suite 300, Broomfield, CO 80021, USA, E-Mail: jwolfe05@harris.com

[2] Harris Geospatial Solutions GmbH, Talhofstraße 32a, D-82205 Gilching, E-Mail: [Thomas.Bahr, Nicolai.Holzer]@harris.com

2 Methodology

A case study was designed to test the effectiveness of ensemble-based machine learning classifiers in discriminating among land-cover classes that were spectrally similar.

2.1 Study Area

The study area is in eastern Texas, USA. The site was chosen because it contains a mix of land cover classes that are predominantly vegetative such as wetlands, forest, agriculture, and rangeland. The area of interest is bounded by coordinates 30.96 – 31.04°N, 96.74 – 96.84°W (Figure 1).

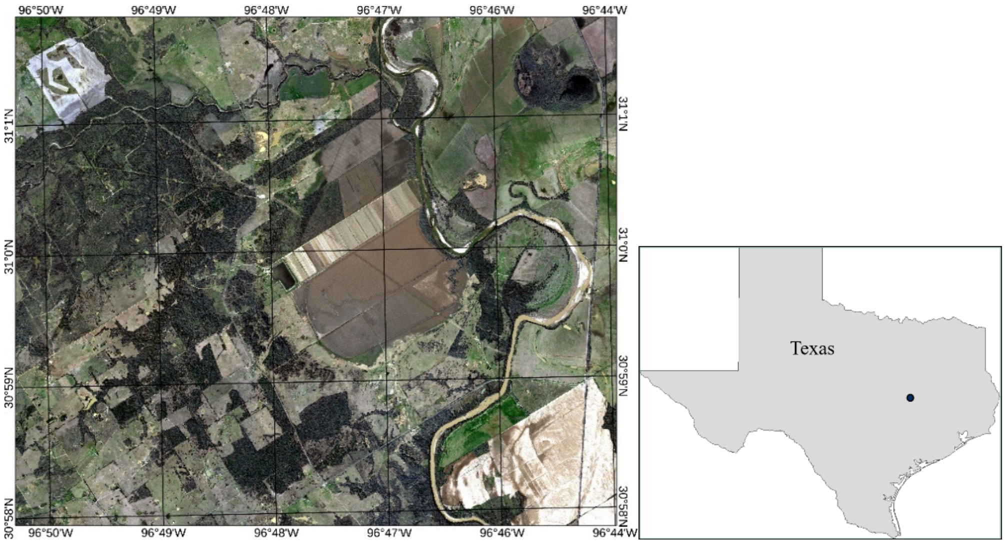

Fig. 1: U.S. Geological Survey orthophoto mosaic of study area (left) and overview map (right) (WOLFE et al. 2018).

2.2 Data Processing

A Sentinel-2A Level-1C image of the study area was used for image classification. The image was acquired on 8 August 2017. Level-1C images are calibrated to top-of-atmosphere reflectance (ESA 2015a). The following processing steps were used with ENVI® 5.5 software to prepare the image for analysis:

- Creating a layer stack of the visible and near-infrared bands, while resampling the 20-meter bands to 10 meters.
- Defining a spatial subset around the area of interest.
- Using the Quick Atmospheric Correction (QUAC) tool to correct the image for atmospheric effects. Pixels represent surface reflectance.
- Creating and applying a mask of all non-vegetation features such as roads, water, and buildings.

The resulting dataset contained 10 spectral bands, listed in Table 1.

Tab. 1: Sentinel-2A spectral bands and associated central wavelengths (ESA 2015b).

Band	Wavelength (nm)	Band	Wavelength (nm)
B2 – Blue	492.4	B6 – Vegetation Red Edge	740.5
B3 – Green	559.8	B7 – Vegetation Red Edge	782.8
B4 – Red	664.6	B8A – Narrow NIR	864.7
B8 – NIR	835.1	B11 – SWIR	1613.7
B5 – Vegetation Red Edge	704.1	B12 – SWIR	2202.4

Field studies to validate ground truth were not feasible, so training data consisted of samples of image pixels that belonged to five feature types: “Nonforested Wetland”, “Rangeland”, “Forest”, “Grass / Pasture”, and “Cropland” (Figure 2). Distinguishing between these land-use types using remote sensing imagery can be challenging since many areas contain a mix of land-use types. In seldom cases the categories are defined by abrupt boundaries, except for possibly “Cropland”.

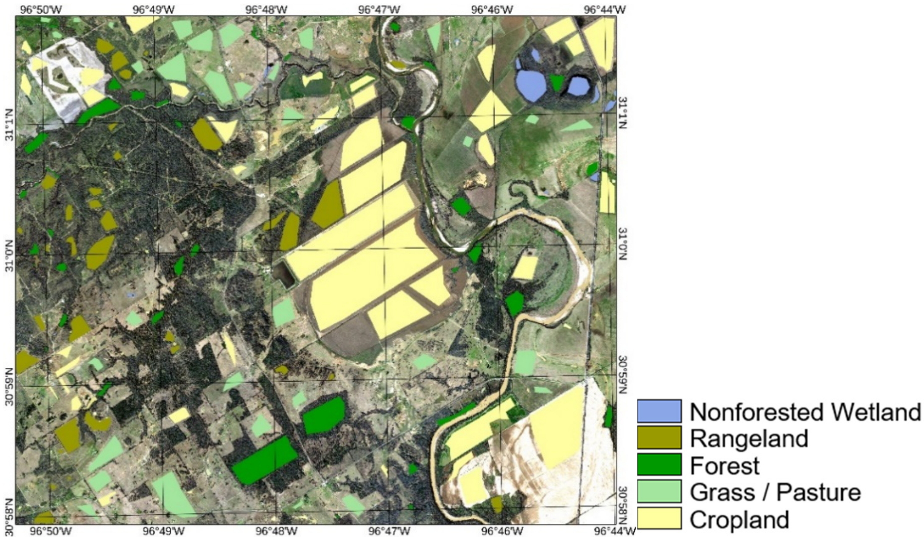

Fig. 2: U.S. Geological Survey orthophoto mosaic of study area with regions of interest (ROI) training samples.

Data-processing steps are described more detailed in WOLFE et al. (2018).

2.3 Machine Learning Classifiers

Machine learning algorithms offer a relatively accurate and efficient way to perform image classification. They can be trained using an initial image and training data, and then used to classify other similar images. For this study, three supervised classifiers were selected from the Python “Scikit-learn” package (PEDREGOSA et al. 2011). These are ensemble learning methods, which train individual decision trees and combine the results to yield the most accurate prediction (SAINI & CHOSH 2017).

The AdaBoost algorithm (FREUND & SCHAPIRE 1995; ZHU et al. 2009) fits a classifier on an original dataset, then fits more copies of the classifier on the same dataset using adjusted weights. This focuses the training on the incorrectly classified pixels.

The Random Forest algorithm has only recently been applied in land-use classification (e. g. KULKARNI & LOWE 2016). It is a decision tree-based classifier whereby several classifiers are trained

and their results are combined based on the votes given by each tree. Random Forest is effective at estimating the importance of individual variables in multi-source classification.
The Extra Trees, or extremely randomized, algorithm (GEURTS et al. 2006) is a decision tree-based classifier. When looking for the optimal split to separate samples of a node into two groups, it draws random splits for each of the randomly selected features. It then chooses the best split. This method is known for its computational efficiency. It is a more randomized version of Random Forest.
These classifiers have significantly less user-defined parameters than traditional machine learning classifiers such as Support Vector Machine (SVM). This makes them ideal candidates for users who are new to remote sensing.
The Python algorithms were imported into IDL® 8.7 and converted to ENVITasks so that the classifiers could run within the ENVI® environment.

2.4 Classification Models

Five classification models were built using the ENVI® 5.5 Modeler, each using a different classifier. These models were simple to construct as they all had the same basic elements (except for the classifier). Figure 3 shows an example of the Extra Trees model. The yellow nodes represent ENVITask-based operations such as extracting training data, defining the classifier parameters, classifying the input image, and exporting the accuracy results to a JSON file on disk. Similar models were built for the Random Forest and AdaBoost classifiers.

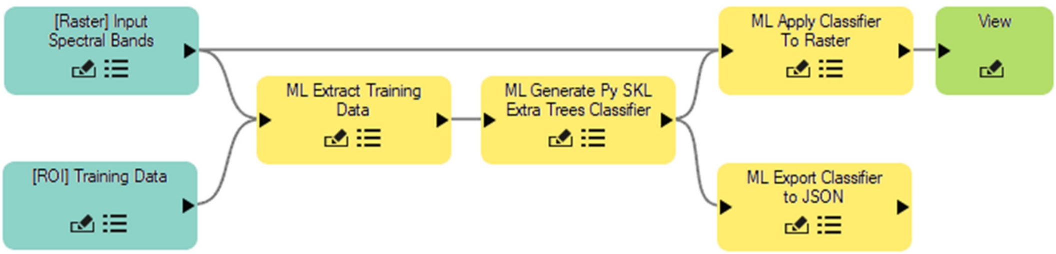

Fig. 3: Extra Trees classification model.

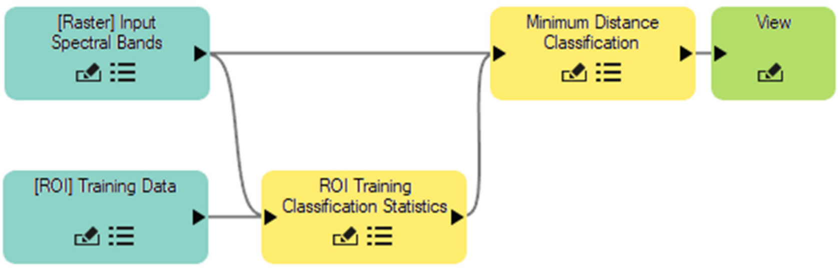

Fig. 4: Minimum Distance classification model.

When building models, users must define the parameters for each classifier. The following values were used for the decision tree-based classifiers:
- Maximum Nodes: 75
- Number of Trees (Extra Trees and Random Forest): 125
- Number of Estimators (AdaBoost): 100

The Maximum Likelihood and Minimum Distance (Figure 4) classifiers did not require any user-defined parameters other than the input image and training ROIs.

3 Results and Discussions

Figures 5-9 show the classification images produced from each model. Tables 2-6 show the resulting accuracy statistics.

The Extra Trees and Random Forest classifiers yielded overall accuracies of 100% for 145,514 training samples. This high accuracy is not uncommon for ensemble-based machine learning classifiers, particularly with an increased number of nodes (75) and trees (125).

The AdaBoost machine learning classifier yielded an overall accuracy of 93% for 210,235 samples.

In general, these three classifiers effectively discriminated between the five mixed-vegetation land-use types.

The two traditional classifiers – Maximum Likelihood and Minimum Distance – yielded accuracies of 84% and 45% for 147,601 samples.

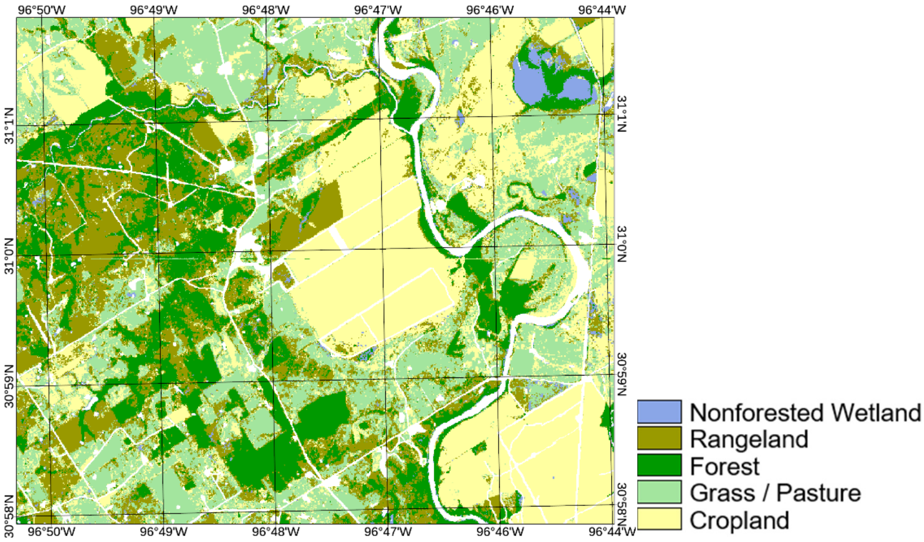

Fig. 5 & Tab. 2: AdaBoost classification image and classifier accuracy metrics.

AdaBoost	Wetland	Rangeland	Forest	Grass / Pasture	Cropland
Producer Accuracy	0.98	0.93	0.81	0.88	0.97
User Accuracy	1.00	0.92	0.92	0.87	0.95
Error of Commission	0.00	0.08	0.08	0.13	0.05
Error of Omission	0.02	0.07	0.19	0.11	0.02
F1	0.99	0.92	0.86	0.87	0.96
Recall	0.98	0.93	0.81	0.88	0.97
Overall accuracy: 0.93			**Kappa coefficient:** 0.89		

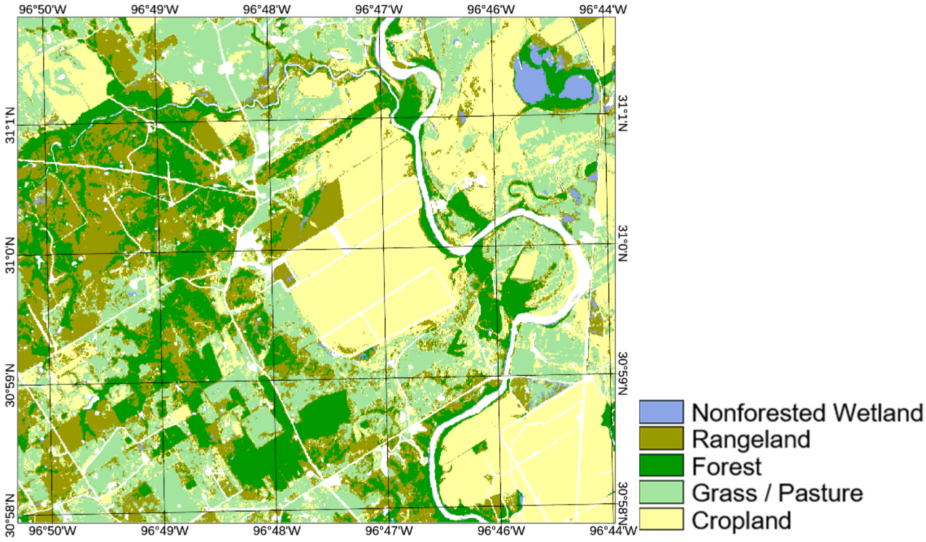

Fig. 6 & Tab. 3: Random Forest classification image and classifier accuracy metrics.

Random Forest	Wetland	Rangeland	Forest	Grass / Pasture	Cropland
Producer Accuracy	1.00	1.00	1.00	1.00	1.00
User Accuracy	1.00	1.00	1.00	1.00	1.00
Error of Commission	0.00	0.00	0.00	0.00	0.00
Error of Omission	0.00	0.00	0.00	0.00	0.00
F1	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00
Overall accuracy: 1.00			**Kappa coefficient:** 1.00		

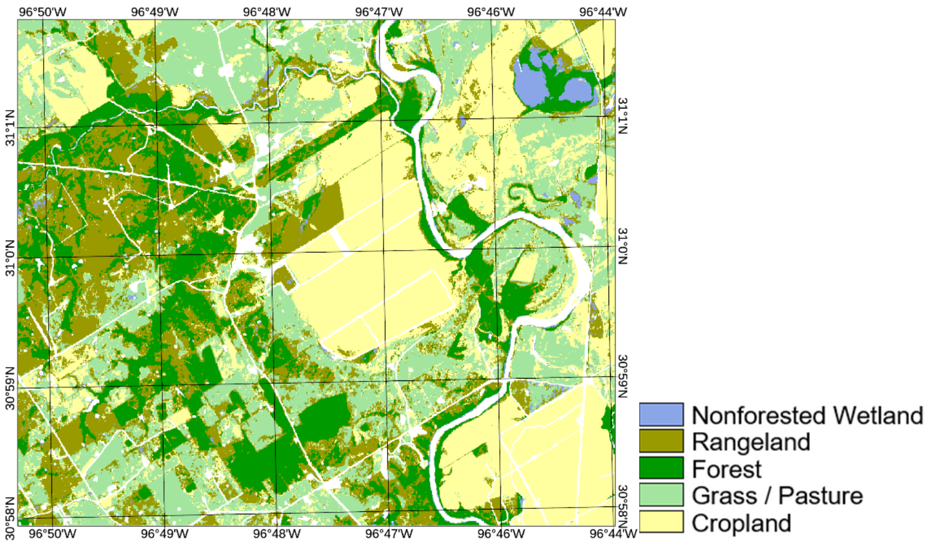

Fig. 7 & Tab. 4: Extra Trees classification image and classifier accuracy metrics.

Extra Trees	Wetland	Rangeland	Forest	Grass / Pasture	Cropland
Producer Accuracy	1.00	1.00	1.00	1.00	1.00
User Accuracy	1.00	1.00	1.00	1.00	1.00
Error of Commission	0.00	0.00	0.00	0.00	0.00
Error of Omission	0.00	0.00	0.00	0.00	0.00
F1	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00
Overall accuracy: 1.00			**Kappa coefficient:** 1.00		

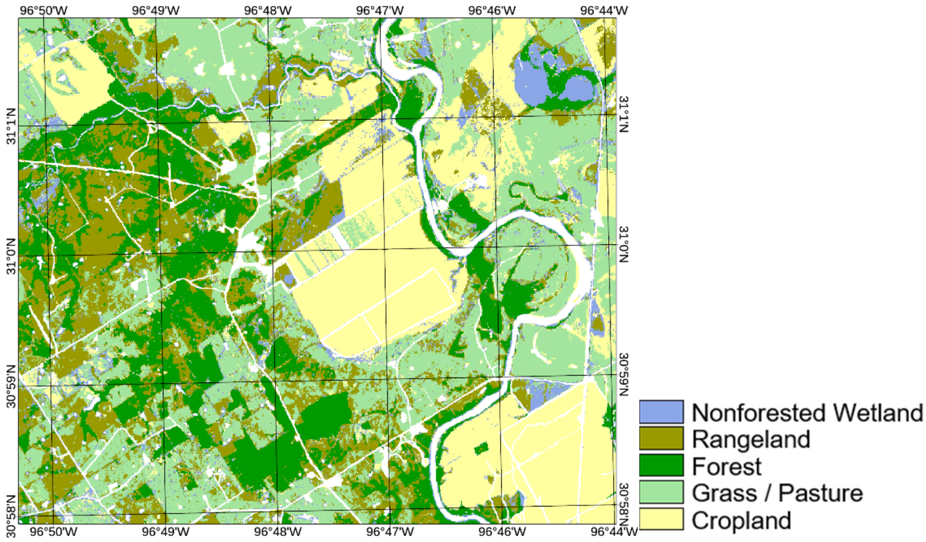

Fig. 8 & Tab. 5: Maximum Likelihood classification image and classifier accuracy metrics.

Maximum Likelihood	Wetland	Rangeland	Forest	Grass / Pasture	Cropland
Producer Accuracy	0.83	0.79	0.93	0.84	0.83
User Accuracy	0.51	0.67	0.94	0.68	0.98
Error of Commission	0.49	0.33	0.06	0.32	0.02
Error of Omission	0.17	0.21	0.06	0.16	0.17
Overall accuracy: 0.84			**Kappa coefficient:** 0.76		

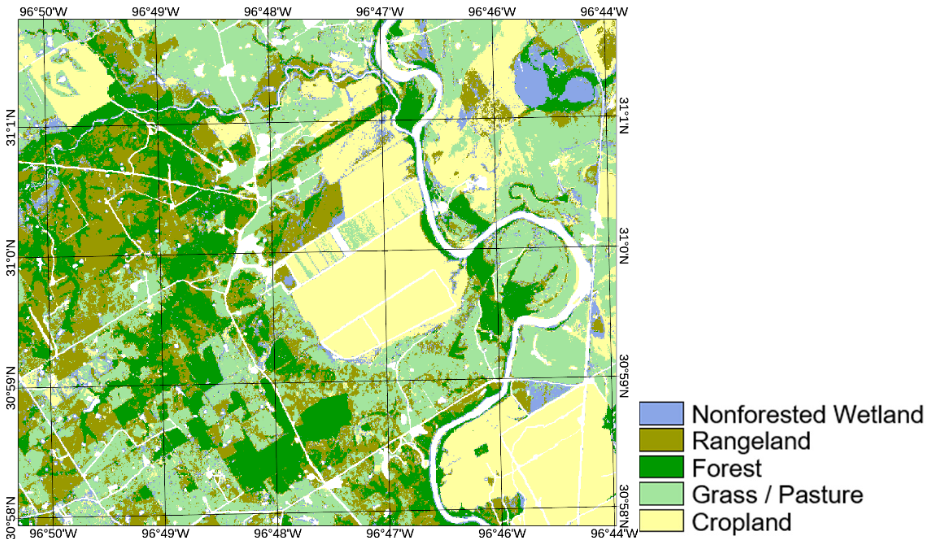

Fig. 9 & Tab. 6: Minimum Distance classification image and classifier accuracy metrics.

Minimum Distance	Wetland	Rangeland	Forest	Grass / Pasture	Cropland
Producer Accuracy	0.50	0.73	0.77	0.57	0.28
User Accuracy	0.26	0.25	0.88	0.40	0.69
Error of Commission	0.73	0.75	0.12	0.60	0.31
Error of Omission	0.50	0.27	0.23	0.43	0.72
Overall accuracy: 0.45			**Kappa coefficient:** 0.28		

4 Conclusions

This research demonstrated how ensemble-based machine learning classifiers were more effective at discriminating among similar land-use classes than traditional classifiers. Python machine learning tools were imported into IDL and converted to ENVITasks. After this initial setup, the ENVITasks could be used in the ENVI® Modeler. The use of models made the process of testing classifiers and comparing results much more efficient, compared to writing API code or invoking tools through a user interface. Model nodes can be easily interchanged to accommodate different trainers and classifiers. Models can also be packaged and deployed to desktop and cloud-computing environments for reuse and further customization. As more machine learning classifiers are implemented in ENVI® software in the future (without the need for a Python-to-IDL interface), the opportunity for building and testing classification models will be simpler and more streamlined.

5 References

EUROPEAN SPACE AGENCY (ESA), 2015a: Sentinel-2 MSI Technical Guide, Level-1C Algorithm. https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-1c/algorithm.

EUROPEAN SPACE AGENCY (ESA), 2015b: Sentinel-2 User Handbook. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook.

FREUND, Y. & SCHAPIRE, R., 1995: A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. Computational Learning Theory, EuroCOLT 1995, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), **904**, VITÁNYI P. (ed), Springer, Berlin/Heidelberg, 23-37.

GEURTS, P., ERNST, D. & WEHENKEL, L., 2006: Extremely Randomized Trees. Machine Learning, **63**(1), 3-42.

GREEN, W., 2012: Econometric Analysis. 7th Edition, Pearson Education, Boston.

KULKARNI, A. & LOWE, B., 2016: Random Forest Algorithm for Land Cover Classification. International Journal on Recent and Innovation Trends in Computing and Communication, **4**(3), 58-63.

PEDREGOSA et al., 2011: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

SAINI, R. & GHOSH, S.K., 2017: Ensemble Classifiers in Remote Sensing: A Review. 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 1148-1152.

WOLFE, J., BAHR, T. & HOLZER, N., 2018: Creating Models of Custom Image Classification Workflows Using Softmax Regression and Support Vector Machine. Journal for Geographic Information Science, 2-2018, JEKEL, T., CAR, A., STROBL, J. & GRIESEBNER, G. (Eds.): GI_Forum 2018, **6**, 240-260.

WOLFE, J., JIN, X., T. BAHR & HOLZER, N., 2017: Application of Softmax Regression and its Validation for Spectral-Based Land Cover Mapping. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., **42**(1/W1), 455-459.

ZHU, J., ZOU, H., ROSSET, S. & HASTIE, T., 2009: Multi-class AdaBoost. Statistics and its Interface, **2**, 349-360.