Comparing Machine Learning Algorithms for Land Cover Classification for Vallon de Nant, Switzerland

Julia Walker [1]

Abstract

This paper aims to compare Random Forest (RF) and Support Vector Machine (SVM) machine learning (ML) classification algorithms in their accuracy for predicting land cover classes for Vallon de Nant, Switzerland. Satellite data for the study area was obtained from Planet Scope to calculate land cover classes. The accuracy metrics of the two algorithms were compared before and after hyperparameter searches were conducted. Conclusions from this study determine that further research into other ML algorithms, such as deep learning algorithms, could improve outcomes.

1. Introduction

Land cover properties are important for various types of analyses in environmental science research. However, detailed land cover data can be difficult to obtain. To combat this data limitation, estimates of land cover can be made based on certain predictors and using machine learning (ML) algorithms. For example, using the Red, Green and Blue spectral bands of satellite imagery can be used to classify land cover, based on the different values of the bands (Gislason et al., 2006). The two ML classification algorithms that will be explored in this paper are Random Forest (RF) and Support Vector Machine (SVM) machine learning (ML). RF is an algorithm based on decision trees and SVM is based on either linear or polynomial separations in order to perform classification (Geron, 2017).

RF has been shown to yield high accuracy results when applied to land cover classification problems (Gislason et al., 2006; Rodriguez-Galiano et al., 2012). There are few articles using SVM classification for land cover classification. Therefore, this is a research gap this report aims to address.

The study area for this land cover classification exercise is Vallon de Nant catchment, Switzerland. It is a natural reserve so there is no urban development within the catchment. Therefore, the only land cover classes are natural. The land cover classes available from the Federal Office of Topography are limited both in classes and spatial coverage (Figure 1). This makes this dataset not ideal to use in subsequent analyses. Therefore, using ML to predict land cover classes for the entire study area is a way to fill this information gap.

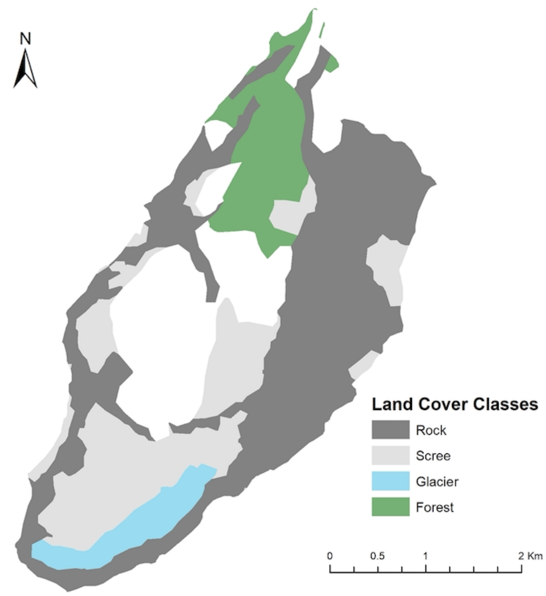

Figure 1. Land cover classes available for Vallon de Nant, Switzerland from the Swiss Federal Office of Topography. Spatial Reference CH1903. Source: map.geo.admin.ch. Map created in ArcMap 10.0 by Julia Walker on November 30, 2022.

2. Methods

2.1. Imagery Data

Satellite imagery of the Vallon de Nant region was obtained from the Planet Scope Scene satellite for one image dating to June 11, 2022 (Figure 2). The image has a spatial resolution of 3m x 3m. The image contained the Red, Green, Blue and Near Infrared (NIR) spectral bands. This image was used as the dataset for training, testing and validation of the two machine learning algorithms (Figure 2). Despite being not open source data, Planet Scope data was chosen to be used as it had a more precise spatial resolution than open source data such as LANDSAT.

2.2. Classified Dataset

Training data was obtained by supervised classification of 30, 362 pixels of the satellite image in ArcGIS 10.0 to generate labelled training data in the form of polygons (Figure 2). Certain pixels were classified as different land cover classes using the classification tool to generate five different classes: low vegetation, glacier, bare rock, forest and sediments. (Figure 2). The classified pixels were used as training data to test the accuracy of RF and SVM in predicting land cover classes. Each class was associated with different spectral band values which are the basis for the prediction of land cover (Gislason et al., 2006).

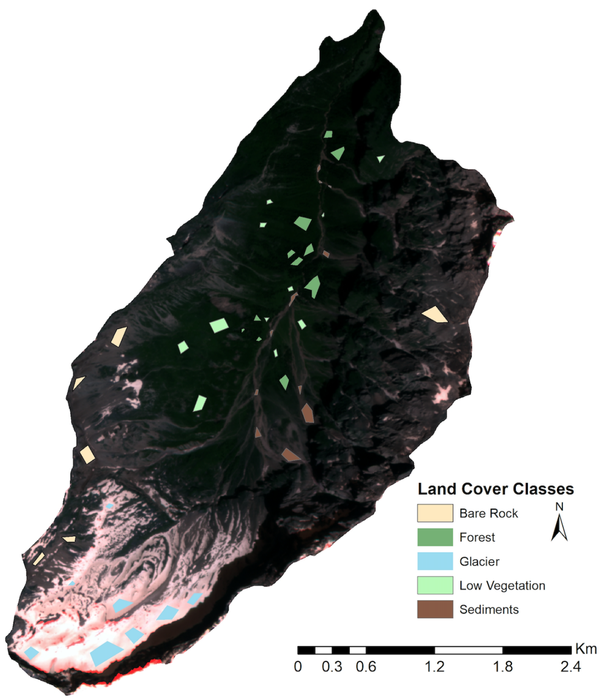

Figure 2. Supervised classification of areas within Vallon de Nant, Switzerland to be used as training data. Imagery source: Planet Scope. Spatial Reference: WGS 1984 UTM Zone 32N. Map created in ArcMap 10.0 by Julia Walker on December 10, 2022 .

2.3. Machine Learning Algorithm Implementation

Python 3.5 in Google Collab was used to implement ML algorithms and performance metrics. The classified dataset was split into training (0.75), testing (0.10) and validation (0.15) datasets in Google Collab. RF and SVM ML algorithms were applied to predict land cover classes using the default parameters for each algorithm (Appendix I). Then a hyperparameter search was conducted to find the best parameters for each algorithm. Finally, performance metrics, such as, accuracy scores, confusion matrices were generated to assess the performance of each of the two algorithms, before and after the best parameters were applied (Appendix I).

2.4. Hyperparameter Search

Hyperparameters control differnt parts of ML algorithms (Geron, 2017). There are many different hyperparameters within ML algorithms. In order to improve the accuracy of each of the two algorithms, GridSearch was used on each of the algorithms to return the best parameter for the model. The number of trees and the maximum features the RF algorithm uses were the hyperparameters chosen for the GridSearch. Additionally, for the SVM algorithm the hyperparameter C, which regularises the model, was chosen to be searched (Geron, 2017).

Table 1. Hyperparameters chosen (number of trees and maximum features) to tune using Grid Search for RF algorithm

	NUM TREES	MAX FEATURES
HYPERPARAMETERS	50 TO 500	2 TO 4
BEST	300	3

Table 2. Hyperparameter chosen (C) to tune using Grid Search for SVM algorithm.

	C
HYPERPARAMETERS	0.1, 1, 10, 100, 1000
BEST	1000

3. Results

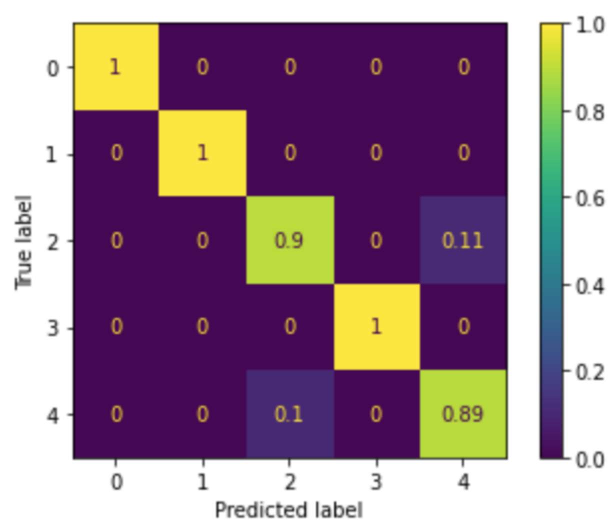

Figure 3. Confusion matrix for RF algorithm based on test data with default hyperparameters. 0= Low Vegetation; 1= Glacier; 2= Bare Rock; 3= Forest; 4= Sediments

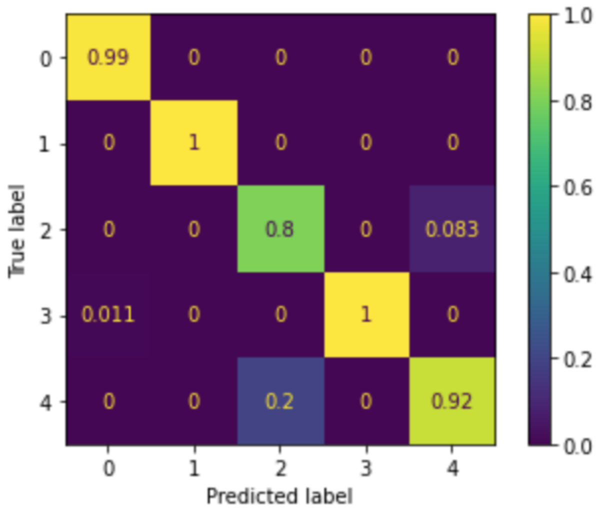

Figure 4. Confusion matrix for SVM algorithm based on test data with default hyperparameters. 0= Low Vegetation; 1= Glacier; 2= Bare Rock; 3= Forest; 4= Sediments

Table 3. Accuracy scores for test and validation datasets before hyperparameter tuning (default) and after tuning (best params) for RF

	DEFAULT	BEST PARAMS
TEST	96.67	96.58
VALIDATION	96.75	96.64

Table 4. Accuracy scores for test and validation datasets before hyperparameter tuning (default) and after tuning (best params) for SVM

	DEFAULT	BEST PARAMS
TEST	94.83	96.58
VALIDATION	95.23	96.64

The results from applying the ML algorithms to the data set showed that the RF algorithm had slightly higher accuracy than the SVM algorithm when using the default hyperparameters for both the validation and test sets (Table 1). This could be why more literature is surrounded using RF as a predictive method for land cover classification over SVM.

However, when tuning each of the hyperparameters for the respective algorithms resulted in the accuracy scores increasing for SVM but decreasing for RF when using the best parameters for both the validation and test sets (Table 2). Overall, the accuracy scores are always above 90 percent, which shows that the algorithms are mostly accurate in predicting land cover class.

Confusion matrices were also produced for the testing set (with the default hyperparameters) to see how well each class was estimated (Figure 3). The confusion matrices show that the two algorithms differ in how accurate they are in predicting the different classes. For example, the RF algorithm was better at predicting the Bare Rock class than SVM but SVM was better at predicting the Sediments class than RF (Figure 3 4).

4. Discussion

Although both ML algorithms produced accuracy scores above 90 percent, they could be improved by using even better algorithms specified to image analysis, such as convolutional neural networks (CNN). This is because CNN learn the dataset by assigning classes based on the spatial features of the image (Kattenborn et al., 2021). It has been studied that CNN are especially beneficial to use for images that contain vegetation (Kattenborn et al., 2021).

The results should also be cautioned as the dataset split was generated randomly, which is not the most ideal way to split spatial data. This is because spatial autorcorrelation can occur, making the model subject to overfitting if the validation and testing sets are spatially close to the testing data (Tonini et al., 2021). Therefore, this study could be improved by implementing a spatially independent split and analysing results.

The reason that the ”best” hyperparameters actually did not yield the best accuracy metric could be due to other hyperparamters that were not included in the Grid Search having more influence on the algorithm’s output. Additionally, maybe the best hyperparameter was out of the range that was tested for.

5. Conclusion

This paper showed that RF and SVM ML algorithms can both be used to accurately predict land cover classes from labeled satellite imagery, with similar accuracy scores. Future work could include exploring more advanced ML algorithms such as convolutional neural networks to improve accuracy outcomes and producing land cover maps.

6. Acknowledgements

Thank you to Marj Tonini, Bryce Finch, Milton Gomez, Jinyan Yu, Emmanuel Emezina, and Professor Tom Beucler for their help and guidance in making this project successful. Finally, thank you to GitHub user Chris Holden for providing tutorials and sample code on using ML for land cover classification.

7. References

Geron, Aurelien. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media, Sebastopol, United States.

Gislason, P. O., Benediktsson, J. A., Sveinsson, J. R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., Fiorucci, P. (2020). A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. Geosciences, 10(3), 105. https://doi.org/10.3390/geosciences10030105

8. Appendix

Access to Planet Scope Scene imagery requires academic institution affiliation. Therefore, the data set cannot be linked here. For more information or to get access to Planet Scope data search here

However, the training data polygons have been made available here on GitHub

Project code is also available here on GitHub