The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification [version 1; peer review: 1 approved]

Fatwa Ramdani1, Muhammad Tanzil Furqon2

1International Public Policy, University of Tsukuba, Tsukuba, 305-8571, Japan
2Geoinformatics Research Group, Informatics Engineering, Brawijaya University, Malang, 65145, Indonesia

Abstract

Background: The availability of urban forest is under serious threat, especially in developing countries where urbanization is taking place rapidly. Meanwhile, there are many classifier algorithms available to monitor the extent of the urban forest. However, we need to assess the performance of each classifier to understand its complexity and accuracy.

Methods: This study proposes a novel procedure using R language with RStudio software to assess four different classifiers based on different numbers of training datasets to classify the urban forest within the campus environment. The normalized difference vegetation indices (NDVI) were then employed to compare the accuracy of each classifier.

Results: This study found that the Extreme Gradient Boosting (XGBoost) classifier outperformed the other three classifiers, with an RMSE value of 1.56. While the Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM) were in second, third, and fourth place with RMSE values of 4.33, 6.81, and 7.45 respectively.

Conclusions: The XGBoost algorithm is the most suitable for urban forest classification with limited data training. This study is easy to reproduce since the code is available and open to the public.

Keywords
xgboost, random forest, support vector machine, neural network, urban forest, classification
This article is included in the Japan Institutional Gateway gateway.

**Corresponding author:** Fatwa Ramdani (fatwa.ramdani.gw@u.tsukuba.ac.jp)

**Author roles:** Ramdani F: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Furqon MT: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The first author, FR, thanks the University of Tsukuba Gateway (F1000) Article Submission Support Program. The second author, MTF, thanks the DIPA program of the Faculty of Computer Science, Universitas Brawijaya for supporting this research. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2022 Ramdani F and Furqon MT. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Ramdani F and Furqon MT. The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification [version 1; peer review: 1 approved] F1000Research 2022, 11:1069 https://doi.org/10.12688/f1000research.124604.1

**First published:** 20 Sep 2022, 11:1069 https://doi.org/10.12688/f1000research.124604.1
Introduction

Trees and vegetation within urbanized areas (buildings, streets, parks, derelict corners, etc) are known as urban forests (http://www.fao.org/). According to the Canadian Urban Forest Strategy (CUFS), urban forest consists of trees, forests, greenspace and related abiotic, biotic and cultural components in areas extending from the urban core to the urban-rural fringe (www.treecanada.ca). The concept of the urban forest is not a new concept, but it has grown in importance, especially in developing countries where urbanization is taking place rapidly.

The availability of trees and vegetation in urbanized areas is important, not only for aesthetic reasons but also for a healthy environment as well as to tackle urban pollution (Brack, 2002; McPherson et al., 2005; Tyrväinen et al., 2005) and microclimate regulator (Ramdani & Setiani, 2014). Furthermore, the access and quality of urban vegetation can increase physical activity as well as residents’ health (Schipperijn et al., 2013; van Dillen et al., 2012).

This urban forest also provides socio-economic benefits and uses. From a social point of view, urban forest creates recreation opportunities, improves inhabitants’ home and work environments, as well as providing a positive impact on physical and mental health (Greenevegen et al., 2006; Ramdani, 2013). From an economic point of view, urban forest could increase property values as well as tourism (Tyrväinen et al., 2005).

Up-to-date information on the presence of urban forest is needed since it gives urban inhabitants a chance to interact with nature. This can have a significant impact on their quality of life in terms of their emotions, bodies, and spirits (Greber et al., 2013). This information could be generated from geospatial datasets, especially raster-based data. The availability of very high-resolution satellite data provides benefits for this task. However, to monitor and map the presence of urban forest we need to evaluate the best method of classifier algorithm.

According to Nguyen et al. (2019), the XGBoost algorithm has some strong and weak points. The strong points include high execution speed and model performance, parallelization of tree construction using all CPU cores during training, distributed computing for training very large models using a cluster of machines, out-of-core computing for very large datasets that do not fit into memory, and cache optimization of data structures and algorithms to make the best use of hardware. However, it is a boosting library that is designed for tabular data, therefore it will not work for other tasks such as natural language processing (NLP).

Previous studies

Some researchers have evaluated the performance of the XGBoost classifier algorithm for satellite image classification as well as other geospatial datasets (Balzotti et al., 2020; Lin et al., 2020; Zheng et al., 2019), Georganos et al. (2018) examined the performance of XGBoost compared with Random Forest (RF) and Support Vector Machine (SVM) for classification of WorldView-3 images, Pleiades images, and aerial photogrammetry of study areas in Burkina Faso, Senegal, and Germany, respectively. They found that the XGBoost classifier algorithm outperformed the RF and SVM algorithms, especially in larger sample sizes. Xu, Ho, et al. (2018) estimated monthly concentrations of ground-level PM2.5 using Moderate Resolution Imaging Spectroradiometer (MODIS) data (https://modis.gsfc.nasa.gov/) and eight different classifier algorithms such as Cubist, RF, and XG Boost. They found that these three classifier algorithms produced better performance than the other classifiers. Another study by Xu, Knudby, et al., (2018) used ten different classifiers to map ambient light at night (ALN) for urban environment studies. One of the classifiers was the XGBoost classifier algorithm. The result showed that XGBoost produced a lower mean absolute error (MAE).

A study by Man et al. (2018) introduced a classification procedure using Landsat 8 images over Hanoi, Vietnam. They compared multiple classifiers such as XGBoost, Logistic Regression, SVM with Radial Basis Function (RBF) kernel, SVM with linear kernel, and Multi-Layer Perceptron (MLP). The study concluded that all classifiers produced high accuracy. Another study by Abdil (2019) also concluded that XGBoost produced high overall accuracy. His study compared SVM with XGBoost, RF, and Deep Learning (DL) classifiers. The results of his research showed that DL was in last place in terms of land cover land use classification with only 73% accuracy using Sentinel-2 images of Sweden and the Baltic region.

Furthermore, there are some studies introducing the application of SVM in remote sensing data classification (Khosravi & Mohammad-Beigi, 2014; Liu et al., 2017; Maulik & Chakraborty, 2012). For instance, Ramdani (2018) introduced a novel procedure to extract oil palm plantation data using Sentinel-2 images. His study compared four different classifiers, which were RF, SVM, K-Nearest Neighbor (KNN), and Gaussian Mixture Model (GMM). He found that the object-based geospatial data feature extraction outperformed the four different classifiers, while the SVM was in second place, followed by KNN, RF, and GMM. Dong et al. (2020) tested a method based on the fusion of an RF classifier and
Convolutional Neural Network (CNN) for a very high-resolution remote sensing (VHRRS) based forest mapping. The study demonstrated the RF classifier produced better results and involved less programming effort.

In terms of the different number of data training, Ramdani et al., (2019) introduced the effect of the different number of data training on the ultra-high resolution of aerial orthomosaic photos derived from an unmanned aerial vehicle. The study concluded that the higher number of data training does not always result in higher accuracy of land use land cover classification. The study compared Multi-Layer Perceptron (MLP) and Radial Basis Function Neural Network (RBFNN).

Although data-driven based-on satellite imagery study is well established for classification applications, its effect on the number of data training on the classification result used in urban forest classification has not been extensively studied. Furthermore, it is challenging to follow and replicate the results of the previous studies. Therefore, the objective of this study was to evaluate the performance of four different classifier algorithms that is XGBoost, RF, SVM, and Artificial Neural Network (ANN) in the accuracy of urban forest classification using different numbers of data training with R language within RStudio 2022.02.3+492 “Prairie Trillium” (RStudio, 2020) as an Integrated Development Environment (IDE). An open-access alternative for RStudio is Jupyter Notebook, which can be run using an Internet connection.

Methods

Study area

The study area was Brawijaya University Campus, located in Malang City, East Java Indonesia (https://ub.ac.id/). With half of the campus covered with trees and vegetation, this is a very suitable area to test the four different classifier algorithms.

According to research by Ramdani et al., (2019), the tree and vegetation of Brawijaya University Campus covers almost 20 ha while the rest is buildings and other infrastructures. This tree and vegetation cover is considered an urban forest by the local government of Malang City. Figure 1 shows the study area of Brawijaya University Campus superimposed with sampling point datasets.

![Figure 1. The study area, Brawijaya University Campus, Malang City, East Java Indonesia.](image-url)
Data and methodology

Data from PlanetScope was collected from https://www.planet.com/ under the Open California Program. Unfortunately, this program has since shut down (https://www.planet.com/). However, researchers are still able to apply for access to the data through the education and science program (https://www.planet.com/science/). Sentinel-2 datasets with 10-meter pixel resolution are also available to the public and can be accessed from https://scihub.copernicus.eu/dhus as an alternative. Regarding the datasets, there are many open-access alternatives to satellite imagery such as Landsat-8 and Landsat-9 with pixel resolutions of 30 m (multispectral), and 15 m (panchromatic), respectively. Landsat-8 and Landsat-9 are available through https://earthexplorer.usgs.gov/. The dataset can be downloaded after registering.

For this study, the acquisition date was September 12, 2019. The Planet imagery has a 3.7 m spatial resolution approximately, with three different bands in the visible spectrum, that is blue with 455–515 nm, green with 500–590 nm, red with 590–670 nm, and a split-frame near Infra-red with 780–860 nm (assets.planet.com/docs/).

The PlanetScope data was clipped using the polygon boundary (see yellow in Figure 1) to minimize the computation time. The boundary was defined based on the outer buildings of the campus. QGIS software version 3.22 “Biatowieza” was used for this step. An open-access alternative for QGIS is GRASS GIS version 7.8.5 or SAGA GIS version 7.9.0.

To classify the remote sensing data, we needed to prepare the training and testing data sets. These data were collected within the urban forest of Brawijaya University campus using a handheld GPS Trimble Juno 3B series. Each land-use type in scenario 1 was represented by five points of training data sets as the minimum number of training data. While scenario 2 used ten points, and scenario 3 used fifteen points of training datasets as the maximum number of training data. There were five land-use types used in this study, which were grass, trees, buildings, roads, and residential.

Training and testing datasets were separated 60:40, which was 60% for training and 40% for testing the result. Training and testing data had the same column which consisted of a class of land use, type of land use, x coordinate, y coordinate, and values of each band that had been extracted using the point sampling tool plugin in QGIS. Table 1 shows the sample of training data sets. The classification process was done in the RStudio environment, while code and datasets used in this study are openly available (Ramdani, 2022; Ramdani & Furqon, 2022).

XGBoost classifier

When working in RStudio, we first made a working directory. In this case, the working directory was “D:/MLinRStudio”. Then we installed and activated the library. There were four libraries needed for the classification using the XGBoost

| PS.1 | PS.2 | PS.3 | class | type | xcoor  | ycoor   |
|------|------|------|-------|------|--------|---------|
| 216  | 225  | 234  | 1     | grass| 677880 | 9120625 |
| 255  | 255  | 255  | 1     | grass| 677895 | 9120609 |
| 250  | 254  | 255  | 1     | grass| 677903 | 9120596 |
| 168  | 165  | 156  | 2     | tree | 677869 | 9121131 |
| 168  | 166  | 150  | 2     | tree | 677637 | 9120667 |
| 164  | 166  | 152  | 2     | tree | 677737 | 9120229 |
| 186  | 183  | 188  | 3     | building| 677948 | 9120950 |
| 255  | 255  | 255  | 3     | building| 678179 | 9120493 |
| 191  | 189  | 190  | 3     | building| 678015 | 9120541 |
| 183  | 178  | 171  | 4     | road | 678060 | 9120890 |
| 183  | 180  | 176  | 4     | road | 677949 | 9120620 |
| 178  | 175  | 172  | 4     | road | 677829 | 9120200 |
| 180  | 178  | 181  | 5     | residential| 678017 | 9121009 |
| 179  | 175  | 179  | 5     | residential| 677594 | 9120937 |
| 190  | 191  | 203  | 5     | residential| 677617 | 9120273 |

Table 1. Sample training datasets. PS.1 to PS.3 are the pixel values of bands of PlanetScope. The “class” column is the code used for classification. The “type” column is land use land cover type. While the xcoor and ycoor columns are coordinate at each point in UTM.
algorithm, that is “raster”, “rtools”, “devtool”, and “xgboost”. According to Friedman (2001), XGBoost is an ensemble tree method that follows the principle of the gradient boosting framework, and uses regularization techniques to control overfitting and model complexity (Chen & Guestrin, 2016).

The original PlanetScope data was a scene with approximate size 24 \times 8 \text{ km}. While the sampling points were collected using handheld GPS Trimble Juno 3B. The next step was to import the clipped dataset of PlanetScope as well as the sampling points data into the RStudio environment. Then we extracted the values of each sample and converted the data frame into a matrix. Finally, we converted the class of sampling point data into a numeric. The classification was then conducted, firstly by training the model and then predicting the result.

To evaluate the classification result we converted the testing data into a spatial object using the X and Y coordinates, then superimposed the testing points on the predicted classification and extracted the values. Finally, the error matrix was produced, and we calculated the classified image.

**Random forest (RF) classifier**

Different from the XGBoost, a larger library is needed to run the RF classifier, that is “raster”, “caret”, “sp” (Bivand et al., 2013), “randomForest”, “rgdal”, and “e1071”. The library “randomForest” was to run the RF classifier algorithm within RStudio. The earlier step was similar, where we needed to set the working directory, install and load the library, and import the raster data from PlanetScope.

The next step was to define the name of the layer of the stack images and load the sampling point dataset. Furthermore, we split the data frame into 60:40 by class and then combined it into single training and a testing data frame. Next was to set up a resampling method in the model training process and then generate the grid search of candidate hyper-parameter values for inclusion in the model training process.

Finally, we ran the RF model and applied it to the dataset. The evaluation method began with the conversion of testing point data into a spatial object using the X and Y coordinates, superimposing it, and extracting the predicted values. The confusion matrix was produced and calculated to evaluate the final result of the classification.

**Support vector machine (SVM) classifier**

There were six libraries needed to run the SVM classifier within the RStudio environment, that is “raster”, “caret”, “sp”, “kernlab” (Karatzoglou et al., 2019), “rgdal”, and “e1071”. The “kernlab” library was specially designated to run the SVM classifier algorithm. SVM is known as an excellent tool for multiclass classification (Hsu & Lin, 2002).

The first step from importing the dataset to setting up the resampling method was similar to the previous RF algorithm. However, in order to generate a grid search of candidate hyper-parameter values for inclusion in the model training process we needed a more complex tuning process to achieve higher accuracy. In the SVM algorithm, we needed to input different parameters to control the non-linearity in the hyperplane and the influence of each support vector.

After the grid search was produced, we then ran the SVM model and applied the model to a dataset. The next step was similar, where we calculated the error matrix and generated the final result.

**Artificial neural network (ANN) classifier**

The difference between the XGBoost, RF, and SVM classifier and ANN are that we needed to adjust the number of neuron units in the hidden layer and adjust the regularization parameter to avoid over-fitting. In this study, we employed 15 neuron units in the hidden layer and 0.1 to 0.5 for the decay parameter to avoid over-fitting. We then ran the ANN model and applied the model to a dataset. The next step was calculating the error matrix and generating the final result.

**Validation using Normalized Difference Vegetation Index (NDVI)**

To calculate the NDVI image we employed equation (1) to the PlanetScope image. The NDVI was first proposed by Tucker (1979). The NDVI was calculated from the visible red and near-infrared light reflected by vegetation. Healthy vegetation absorbs most of the visible light that hits it, and reflects a large portion of the near-infrared light. Unhealthy or sparse vegetation reflects more visible light and less near-infrared light (Tucker, 1979).

\[
\text{NDVI} = \frac{\text{NIR (band 4)} - \text{Red (band 3)}}{\text{NIR (band 4)} + \text{Red (band 3)}}
\]

Theoretically, the indices should produce values ranging from −1 to +1; however, in our study area NDVI values ranged from −0.11 to 0.5. We then reclassed the NDVI image into five different classes, which were non-vegetation...
(-0.11–0.01), low vegetation (0.02–0.14), light vegetation (0.15–0.27), medium vegetation (0.27–0.4), and high vegetation (>0.41) covered. Then we extracted the highest three classes and combined them into a single class of vegetation. These data were then employed as the testing data for the accuracy assessment of the best scenario. The Root Mean Square Error (RMSE) was used to evaluate the accuracy between the four different classifiers and NDVI. The RMSE compares a predicted value (\(\hat{P}_i\)) of four classifiers (\(n\)) and an observed value (\(O_i\)) of NDVI (Equation 2).

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{P}_i - O_i)^2}{n}}
\]  

**Results**

**Scenario 1: five samples of each class**

The final result of the first scenario is shown in Figure 2. The green colour represents the grass, dark green represents the trees, red represents the buildings, yellow represents the road, and residential is represented by orange.

The RF, SVM, and NN algorithm produced the lowest accuracy level, it can be seen from the Figure 2 that all classes were classified into the building. The accuracy and kappa value could only achieve approximately 2% and 0, respectively.

The XGBoost algorithm produced the highest accuracy level, with an accuracy of 93% and a kappa value of 0.92. Figure 2 shows that tree land-use type dominated, followed by roads, buildings, and grass, while residential was the lowest.

**Scenario 2: ten samples of each class**

In this scenario, the XGBoost algorithm still outperformed the other three classifier algorithms with accuracy of 93% achieved and a kappa value of 0.92. The RF algorithm accuracy followed in second with 91% accuracy and 0.88 kappa value. The NN algorithm was in third position with 83% accuracy and 0.79 kappa value, and the SVM algorithm was in last position with 60% accuracy and 0.49 kappa value.

In this scenario, all three classifier algorithms increased the accuracy and kappa value. However, the XGBoost algorithm still performed better than the other three classifier algorithms. The final result of the scenario 2 classification is shown in Figure 3.

![Figure 2. Classified images of PlanetScope using four different algorithms in scenario 1. Trees are shown in dark green; roads are shown in yellow; buildings are shown in red; grass is shown in light green; and residential areas are represented(279,742),(979,958)
Figure 3. Classified images from PlanetScope using four different algorithms in scenario 2. Trees are represented by dark green; roads are represented by yellow; buildings are represented by red; grass is represented by light green; and residential areas are represented by orange.

Figure 4. Classified images of PlanetScope using four different algorithms in scenario 3. Trees are represented by dark green; roads are represented by yellow; buildings are represented by red; grass is represented by light green; and residential areas are represented by orange.
Scenario 3: fifteen samples of each class
The third scenario produced different values of accuracy and kappa for all four classifier algorithms. The XGBoost algorithm was still in first position, however, the accuracy decreased slightly to 91% and the kappa value decreased to 88. The RF algorithm followed in second place with 77% accuracy, a dramatic decrease from 91% accuracy in the second scenario. The kappa value of the RF algorithm also decreased to 0.71 from 0.88 in the second scenario.

The accuracy of the NN algorithm also decreased from 83% in the second scenario to 65% in the third scenario and the kappa value from 0.79 to 0.56. The SVM algorithm was still in last place with 60% accuracy and 0.49 kappa value, but there were no changes from the second scenario. The final result of the scenario 3 classification is shown in Figure 4.

Validation
The vegetation extracted from the PlanetScope NDVI image was as large as 15.46 ha. The results of the four different classifiers and NDVI are compared and summarized in Table 2 while Figure 5 shows the NDVI image of the study area. Table 3 summarizes the accuracy and kappa values of four different classifiers. Once again, the XGBoost classifier

| Algorithm  | Observed vegetation (NDVI) | Predicted vegetation area (Ha) | Difference | RMSE |
|------------|---------------------------|-------------------------------|------------|------|
| RF         | 15.46                     | 19.81                         | 4.35       | 6.81 |
| SVM        | 15.46                     | 26.81                         | 11.35      | 7.45 |
| ANN        | 15.46                     | 21.38                         | 5.93       | 4.33 |
| XGBoost    | 15.46                     | 17.02                         | 1.56       | 1.56 |

Table 2. The observed value of the NDVI, the predicted values of four classifiers, the difference, and the RMSE value. The “observed vegetation (NDVI)” is a value acquired from PlanetScope data. The “predicted vegetation area (Ha)” was acquired from the four different classifiers. The difference is the difference between the “observed vegetation (NDVI)” and “predicted vegetation area (Ha)”. RF, Random Forest; SVM, Support Vector Machine; ANN, Artificial Neural Network; XGBoost, Extreme Gradient Boosting.

Figure 5. The NDVI image of the study area.
outperformed the other three classifiers with the lowest RMSE value of 1.56. The ANN classifier followed in second place with an RMSE value of 4.33, the RF classifier was in third with an RMSE value of 6.81, and the SVM classifier was in last place with an RMSE value of 7.45.

Further, Figure 6 shows the map comparison between NDVI and the classified result of the four different classifier algorithms. It can be seen that the map of vegetation produced using the XGBoost classifier is similar to the map of vegetation derived from NDVI.

### Discussion

This study shows that the XGBoost classifier algorithm was the most suitable to use for urban classification with very limited data training. These methods can be used by other researchers with limited knowledge of the study area.

However, the proposed method is computationally intensive. In order to reproduce this work a computer with high specifications must be used. Therefore, we recommend using a computer with gaming specifications; a minimum of 8GB of RAM, 4GB of GPU, a NVIDIA graphics card with CUDA, and 500GB of SSD is needed. In this study computational performances, execution time, and complexity of the satellite imagery were not evaluated.
Since the study area is located in tropics, there are no phenological cycles as in temperate regions, where leaf canopy decreases gradually between late summer and early winter (Schuster et al., 2020). We suggest further research is conducted using the proposed method in a study area located outside the tropical zone and multi-temporal analysis should be considered.

In principle all four classifier algorithms need training data sets. However, the ANN classifier uses layers as a basis for computation while RF and XGBoost use loss function to compute the mean decrease, and SVM classifier uses hyper-parameter values. These differences clearly produced different results in this study. Despite these differences, the results of XGBoost produced consistent results while the other three classifiers did not.

Conclusions
This study evaluates the implementation of the XGBoost classifier algorithm with three other classifiers (Random Forest, Support Vector Machine, and Artificial Neural Network). The result demonstrates that the XGBoost classifier algorithm consistently outperformed the other three classifiers. The study found that the different numbers of data training affect the final accuracy as well as the kappa value.

When compared with observed dense vegetation of NDVI, once again the XGBoost algorithm outperformed the other three classifiers with the lowest RMSE value. All three classifier algorithms produced inconsistent accuracy and kappa values when using different numbers of data training. While the XGBoost classifier always produced high accuracy in all scenarios.

This study found that more data training does not always lead to higher or better accuracy with the RF, SVM, and ANN classifiers. However, when using the XGBoost classifier, the small number of data training still produces high accuracy. The novel procedure proposed in this study is reproducible with the availability of the code open to the public. Compared to the other three classifiers, the XGBoost classifier code is the simplest.

Further research could be conducted to compare the XGBoost classifier with other machine learning classifier algorithms such as the Classification and Regression Tree (CART) classifier, or even with the Genetic Evolution (GE) algorithm and Swarm Particle Optimization (SPO) with limited training datasets. Computational performances, execution time, and complexity of the satellite imagery also needs to be evaluated.

Data availability
Underlying data
Mendeley Data: Underlying data for ‘The simplicity of XGBoost algorithm versus the complexity of Random Forest, Support Vector Machine, and Neural Networks algorithms in urban forest classification’, https://doi.org/10.17632/j739yc6cgc.1 (Ramdani & Furqon, 2022).

This project contains the following underlying data:

- Data file 1: PS.tif (PlanetScope raster data)
- Data file 2: samplingPS10.csv (Data for data training and testing)
- Data file 3: samplingPS15.csv (Data for data training and testing)
- Data file 4: samplings.csv (Data for data training and testing)
- Data file 5: ub_ps_names.csv (Data for data training and testing)
- Data file 6: NN_PlanetScope.R
- Data file 7: RF&SVM.R
- Data file 8: xgBoostalgorithm.R

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0)
Software availability
Source code available from: https://github.com/fatwaramdani/f1000
Archived source code at time of publication: https://doi.org/10.5281/zenodo.7014120 (Ramdani, 2022)
License: CC-BY 4.0

Acknowledgments
The authors thank the valuable comments by the reviewers as well as the assistance of the journal editor.

References

Abdi AM: Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GIScience and Remote Sensing. 2019; 57(00):1–20.
Publisher Full Text

Baatz S, Asser GP, Adkins ED, et al.: Spatial drivers of composition and connectivity across endangered tropical dry forests. J. Appl. Ecol. 2020; 57(8):1593–1604.
Publisher Full Text

Bivand R, Pebesma E, Gomez-Rubio V: Applied spatial data analysis with R. Second ed. Springer; 2013.

Brack CL: Pollution mitigation and carbon sequestration by an urban forest. Environ. Pollut. 2002; 116(SUPPL. 1):5195–5200.
Publisher Full Text

Chen T, Guestrin C: XGBoost: A scalable tree boosting system. J. Mach. Learn. Res. 2016; 15(2):520
Publisher Full Text

Chun H, Chen T, Guestrin C: Learning frameworks and libraries for large-scale data mining: a survey. Artif. Intell. Rev. 2019; 52(1):77–114.
Publisher Full Text

Dong L, Xing L, Liu T, et al.: Very High Resolution Remote Sensing Imagery Classification Using a Fusion of Random Forest and Deep Learning Technique-Subtropical Area for Example. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2020; 13:113–128.
Publisher Full Text

Friedman JH: Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 2001; 15(5):41–1232.
Publisher Full Text

Georgoulias S, Grippa T, Vanhuyse S, et al.: Very High Resolution Object-Based Land Use-Land Cover Urban Classification Using Extreme Gradient Boosting. IEEE Geosci. Remote Sens. Lett. 2018; 15(4):607–611.
Publisher Full Text

Grebner E, Betingger P, Sinn JP: Urban Forestry, Introduction to Forestry and Natural Resources. 2013: 385–405.
Publisher Full Text

Groenewegen PP, Van Den Berg AE, De Vries S, et al.: Vitamin G: Effects of green space on health, well-being, and social safety. BMC Public Health. 2006; 6:1–9.
Publisher Full Text

Hsu CW, Lin CJ: A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 2002; 13(2):415–425.
Publisher Full Text

Karatzoglou A, Smola A, Hornik K, et al.: kernlab: Kernel-Based Machine Learning Lab. 2019.
Reference Source

Khosravi I, Mohammad-Beigi M: Multiple Classifier Systems for Hyperspectral Remote Sensing Data Classification. Journal of the Indian Society of Remote Sensing. 2014; 42(2):423–428.
Publisher Full Text

Lin P, Pan M, Allen GH, et al.: Global Estimates of Reach-Level Bankfull River Width Leveraging Big Data Geospatial Analysis. Geophys. Res. Lett. 2020; 47(5):1–12.
Publisher Full Text

Liu P, Cho KJR, Wang L, et al.: SVM or deep learning? A comparative study on remote sensing image classification. Soft. Comput. 2017; 21(12):7053–7065.
Publisher Full Text

Man CD, Huy TT, Bui HQ, et al.: Improvement of land-cover classification over frequently cloud-covered areas using landsat 8 time-series composites and an ensemble of supervised classifiers. Int. J. Remote Sens. 2018; 39(4):1243–1255.
Publisher Full Text

Maulik U, Chakraborty D: A novel semisupervised SVM for pixel classification of remote sensing imagery. Int. J. Mach. Learn. Cybern. 2012; 3(4):247–258.
Publisher Full Text

McPherson G, Simpson JR, Peper P, et al.: Municipal forest benefits and costs in five US cities. J. For. 2005; 103(8):411–416.
Publisher Full Text

Nguyen G, Dlugolinsky S, Bobák M, et al.: Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif. Intell. Rev. 2019; 52(1):77–114.
Publisher Full Text

Ramdani F: Extraction of Urban Vegetation in Highly Dense Urban Environment with Application to Measure Inhabitants’ Satisfaction of Urban Green Space. J. Geogr. Inf. Syst. 2013; 05(April):117–122.
Publisher Full Text

Ramdani F: Recent expansion of oil palm plantation in the most eastern part of Indonesia: feature extraction with polarimetric SAR. Int. J. Remote Sens. 2018; 40(00):7371–7388.
Publisher Full Text

Ramdani F: R script for urban forest extraction using PlanetScope dataset. [Code]. Zenodo. 2022.
Publisher Full Text

Ramdani F, Furqon MT: Urban forest. [Dataset]. Mendeley Data, V1. 2022.
Publisher Full Text

Ramdani F, Furqon MT, Setiawan BD, et al.: Analysis of the application of an advanced classifier algorithm to ultra-high resolution unmanned aerial aircraft imagery – a neural network approach. Int. J. Remote Sens. 2020; 41(9):3266–3286.
Publisher Full Text

Ramdani F, Setiawan BD: Temporal analysis of temperature and relative humidity in Bandung City, Indonesia. Urban Ecosystems. 2014; 17(2):473–487.
Publisher Full Text

RStudio: RStudio. 2020.
Reference Source

Schipperijn J, Bentsen P, Troelsen J, et al.: Associations between physical activity and characteristics of urban green space. Urban Forestry and Urban Greening. 2013; 12(1):109–116.
Publisher Full Text

Schuster MJ, Wragg PD, Williams LJ, et al.: Phenology matters: Extended spring and autumn canopy cover increases biotic resistance of forests to invasion by common buckthorn (Rhamnus cathartica). For. Ecol. Manag. 2020; 464(November 2019):118067.
Publisher Full Text

Tucker CJ: Red and photograhic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 1979; 8(2):127–150.
Publisher Full Text

Tyrénäinen L, Pahlevi, Seelan K, et al.: Benefits and uses of urban forests and trees. Urban Forests and Trees: A Reference Book. 2005: 81–114.
Publisher Full Text

van Dillen SME, de Vries S, Groenewegen PP, et al.: Greenspace in urban neighbourhoods and residents’ health: Adding quality to quantity. J. Epidemiol. Community Health. 2012; 66(6):e8–e9.
PubMed Abstract | Publisher Full Text

Xu Y, Ho HC, Wong MS, et al.: Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly...
Xu Y, Knudby A, Côté-Lussier C: Mapping ambient light at night using field observations and high-resolution remote sensing imagery for studies of urban environments. *Build. Environ.* 2018; 145(August): 104–114.

Zheng Z, Ma Q, Jin S, et al.: Canopy and Terrain Interactions Affecting Snowpack Spatial Patterns in the Sierra Nevada of California. *Water Resour. Res.* 2019; 55(11): 8721–8739.
Mustafa Zeybek
Selcuk Universitesi, Konya, Konya, Turkey

1. Abstract and Introduction Sections: The abstract and introduction sections are appropriate.

2. Methodology Section: The methodology section could be improved to be more streamlined instead of resembling a tutorial. For example, the section on Artificial Neural Networks (ANN) is very short and could be expanded upon. Additionally, it would be beneficial to emphasize the improvements made in your study compared to previous research. Currently, the section primarily focuses on the applications of various methods. Please include details on the calculation of the kappa index.

3. Sample Data Limitations: It would be helpful to discuss any limitations or determine the optimal sample size for your study. Provide insights into the potential constraints or recommendations for future research.

4. Figure 2: Figure 2 appears to be unclear or abnormal. I recommend reviewing the parameters of the algorithm used and ensuring that the figure accurately represents the data and results.

5. Discussion Section: The discussion section needs improvement. Provide more depth and analysis of the findings. Consider addressing the implications of the results, comparing them to existing literature, and discussing any potential limitations or future directions for research.

Overall, the review could be enhanced by addressing these points to provide a more comprehensive and informative paper.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes
Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Landslide, UAV, LiDAR, Classification, remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com