FEATURE SELECTION ON SENTINEL-2 MULTI-SPECTRAL IMAGERY FOR EFFICIENT TREE COVER ESTIMATION

Usman Nazir[1,2]*, Momin Uppal*[1]*, Muhammad Tahir*[1]*, and Zubair Khalid*[1]

[1] Department of Electrical Engineering, Syed Babar Ali School of Science and Engineering
Lahore University of Management Sciences (LUMS), Lahore, Pakistan
{usman.nazir, momin.uppal, tahir, zubair.khalid}@lums.edu.pk
[2] Planetary Health Informatics Lab, University of Oxford
usman.nazir@ndorms.ox.ac.uk

ABSTRACT

This paper proposes a multi-spectral random forest classifier with suitable feature selection and masking for tree cover estimation in urban areas. The key feature of the proposed classifier is filtering out the built-up region using spectral indices followed by random forest classification on the remaining mask with carefully selected features. Using Sentinel-2 satellite imagery, we evaluate the performance of the proposed technique on a specified area (approximately 82 acres) of Lahore University of Management Sciences (LUMS) and demonstrate that our method outperforms a conventional random forest classifier as well as state-of-the-art methods such as European Space Agency (ESA) WorldCover 10m 2020 product as well as a DeepLabv3 deep learning architecture.

Index Terms— Random Forest Classifier, Spectral Indices, Sentinel-2 Satellite, European Space Agency (ESA) WorldCover, DeepLabv3

1. INTRODUCTION

The presence of easily accessible multispectral satellite imagery has expanded the range of potential applications across diverse fields. An important example is automated detection of trees and green spaces that are significant contributors to ecosystem services such as air purification and carbon sequestration. Recent studies include [1] and [2] for global monitoring of environment and forest cover using Sentinel-2 imagery. A Copernicus Sentinel-2B satellite, launched in 2017 provides 13 bands with spatial resolution from 10 m to 60 m. The high spatial and temporal resolution of data from this satellite is specifically designed for vegetation monitoring. For tree cover estimation, a broad range of methodologies have been presented in the literature, e.g., [1, 3, 4, 5]. The authors in [3] proposed a data fusion method of different spatial resolution satellite imagery for forest-type mapping. Forest cover change is mapped in [4] using spatial resolution of 25 m to 30 m. A spatio-temporal study on forest cover change using GIS and remote sensing techniques in Ethiopia valley is presented in [5]. In [1], a simple tree cover (referred to as 'forest cover') approach using three different land cover (LC) products is employed in Finland. Clearly, most of these approaches focus on forest mapping – a gap exists in urban tree cover estimation in developing countries with low resolution imagery.

In this paper, We propose a multi-spectral classifier (that uses a mixture of spectral bands *and* indices) for tree cover estimation in urban areas. The key aspects of the proposed classifier include a masking stage for filtering out built-up areas, followed by a random forest classifier operating on appropriately selected features. For performance evaluation, we manually annotate 3768 trees in Lahore, Pakistan[1]. We demonstrate that on account of suitable feature selection and masking mechanism, our proposed approach applied to low resolution imagery achieves a higher accuracy level compared to that obtained by the European Space Agency (ESA) WorldCover product [6] as well as a more computationally demanding deep learning architecture DeepLabv3 [7] applied on high-resolution imagery.

The subsequent sections of this paper are structured as follows: Section 2 delves into a comprehensive analysis of the methodology, while Section 3 showcases the evaluation results comparing our proposed methodology with state-of-the-art models. Finally, Section 4 concludes the paper.

2. METHODOLOGY

The proposed methodology, illustrated in Fig. 1, consists of four stages. These include 1) Pre-processing, 2) Feature selection, 3) Masking, and finally 4) Random Forest Classification. The details of each stage are provided in the text below.

We acknowledge the support of the Higher Education Commission of Pakistan under grant GCF-521.

[1]This dataset is being made publicly available at `https://city.lums.edu.pk/products/`

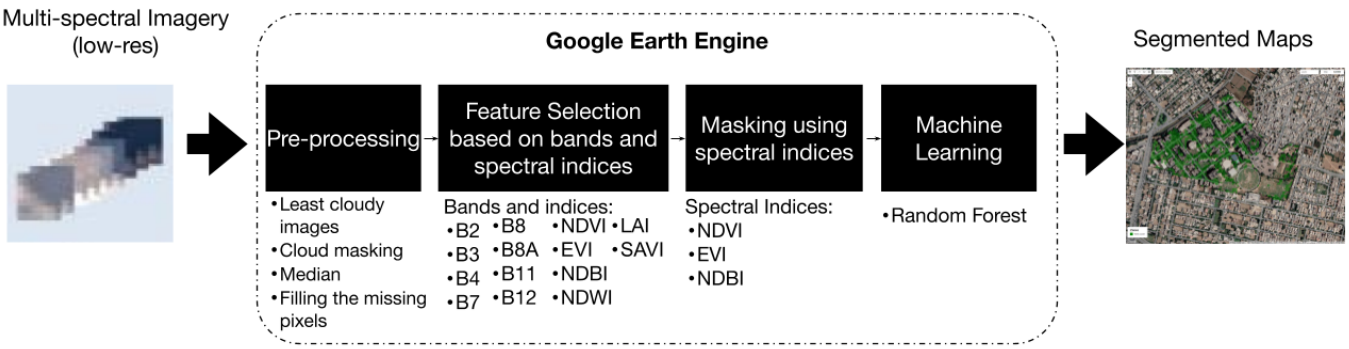

Fig. 1. Proposed methodology for tree cover estimation with feature selection of spectral bands and indices.

Table 1. Tree cover area is predicted using Sentinel-2 imagery and RF classifier by employing various feature selection techniques. The Lahore University of Management Sciences (LUMS) study region spans 82.165 acres of area with 23 acres area of actual tree cover.

ROI	Model	Pred. area (acres)	Masking	Spectral indices	Pixel-wise Test Accuracy (%)	Kappa Score
LUMS	RF-spectral-bands	29.5	No	No	0.93	0.81
LUMS	RF-spectral-indices	28	No	Yes	0.95	0.88
LUMS	**Proposed**	25	Yes	Yes	**0.99**	**0.92**
LUMS	ESA WorldCover Product [6]	16	-	-	0.74	-
LUMS	DeepLabv3 [7]	28	No	No	0.80	-

2.1. Pre-processing

We divide the pre-processing of data into multiple steps. Initially, the images from a multi-spectral satellite containing less than 10% cloud cover for the region of interest (LUMS) are passed through a cloud masking operation that removes cloud cover from these images. Next a median of these images is taken for each month. Finally multiple images are stacked together to generate a single combined image of the region of interest.

2.2. Feature selection

For classification, we included eight bands of Sentinel-2 imagery as the feature set. These include B2 (Blue), B3 (Green), B4 (Red), B7 (Red Edge 3), B8 (NIR), B8A (Red Edge 4), B11 (SWIR 1) and B12 (SWIR 2). In addition, We also chose six spectral indices in our feature set. These include the Normalized Difference Vegetation Index (NDVI) [8], Enhanced Vegetation Index (EVI) [9], Normalized Difference Built-up Index (NDBI) [10], Normalized Difference Moisture Index (NDMI) [11], Leaf Area Index (LAI) [12] and Soil Adjusted Vegetation Index (SAVI) [13]. In general, regions with tree cover typically exhibit high vegetation indices (EVI, NDVI), NDMI, LAI, and SAVI, while showing notably low values for NDBI. Some background about these indicies is given below.

NDVI: This index [8] describes the difference between visible and near-infrared reflectance of vegetation cover and can be used to estimate the density of green on an area of land. This is computed from the the NIR and the Red bands measurements as follows

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \tag{1}$$

EVI and LAI: EVI [9] is similar to NDVI and can be used to quantify greenness of vegetation. However, EVI corrects for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation. It is computed as

$$\text{EVI} = 2.5 \times \frac{\text{NIR} - \text{Red}}{\text{NIR} + 6 \times \text{Red} - 7.5 \times \text{Blue} + 1} \tag{2}$$

On the other hand, LAI [12] is used to estimate crop growth and yield through the following empirical formula

$$\text{LAI} = 3.618 \times \text{EVI} - 0.118 \tag{3}$$

SAVI: This index [13] attempts to minimize soil brightness influences using a soil-brightness correction factor. This is often used in arid regions where vegetative cover is low, and it outputs values between -1 and 1 through the following relationship

$$\text{SAVI} = \frac{0.5 \times (\text{NIR} - \text{Red})}{\text{NIR} + \text{Red} + 0.5} \tag{4}$$

NDWI: This is a satellite-derived index [14] from the NIR and the SWIR channels. The NDWI is used to monitor changes related to water content in water bodies as they

Table 2. Pixel-wise accuracy and Kappa score of proposed model with different feature set on LUMS study region.

(RF + Masking +) Features set	**Pixel-wise Test Accuracy (%)**	Kappa Score
Eight multispectral bands + NDVI	0.96	0.76
Eight multispectral bands + NDVI + NDWI + NDBI + EVI	0.97	0.80
Eight multispectral bands & All spectral indices	**0.99**	**0.92**

Fig. 2. Qualitative results using feature selection on Sentinel-2 multi-spectral imagery for efficient tree cover estimation.

strongly absorb light in visible to infrared electromagnetic spectrum.

$$\text{NDWI} = \frac{\text{NIR} - \text{SWIR1}}{\text{NIR} + \text{SWIR1}} \tag{5}$$

NDBI: This index [15] uses the NIR and SWIR bands to emphasize manufactured built-up areas. It aims to mitigate the effects of terrain illumination differences as well as atmospheric effects.

$$\text{NDBI} = \frac{\text{SWIR} - \text{NIR2}}{\text{SWIR} + \text{NIR2}} \tag{6}$$

2.3. Masking

Masking process involves the following two steps.
Applying the Vegetation Index. The EVI or NDVI values are calculated for each pixel in the satellite imagery. These values indicate the presence and density of vegetation. In this case, a threshold of 0.2 is set, implying that any pixel with an EVI or NDVI value equal to or below 0.2 is considered non-vegetated or sparsely vegetated. Such regions are likely to include built-up areas, as they have less vegetation cover.
Utilizing the Built-up Index. Simultaneously, the NDBI values are computed for each pixel. This index highlights the presence and extent of built-up areas. High positive NDBI values indicate the dominance of built-up surfaces, while low or negative values represent non-built-up or natural areas.

By combining the results of both the vegetation index and built-up index, the filtering process identifies and excludes pixels with low vegetation (pixels for which both EVI and NDVI are less than or equal to 0.2) *and* high built-up signatures (pixels that have positive NDBI values).

2.4. Random Forest (RF) Classification

The masking operation described above aims to retain only the non-built-up or natural regions for input to the classification module. For the purpose, we utilize an RF classifier which is an example of ensemble learning where each model is a decision tree. Ensemble learning creates a stronger model by aggregating the predictions of multiple weak models, such as decision trees in our case. To train the RF classifier, we need to have at least two classes.

We combine multiple sample points along with their corresponding class labels (representing trees or non-trees), divide the samples into an 80% training set and a 20% validation set, train a random forest classifier with the features described above, and then use the trained classifier to classify the input image. In the process of an RF model training, the user defines the number of features at each node in order to generate a tree. The classification of a new dataset is done by passing down each case of the dataset to each of the grown trees, then the forest chooses a class having the most votes of the trees for that case. More details on RF can be found in Breiman [16]. The main motivation behind choosing RF for this study is its ability to efficiently handle large and high dimensional datasets [17, 18].

3. EVALUATION

The proposed methodology is applied to satellite imagery for the year 2021 and its performance compared to other benchmarks is shown in Table 1. As the results indicate, RF with all multi-spectral bands performs better than the ESA WorldCover product [6] and DeepLabv3 [7]. RF with spectral indices achieve higher accuracy as compared to RF with only spectral bands. Finally, the proposed model accomplishes higher pixel-wise accuracy and Kappa score as compared to all other models (see Table 2).

Results indicating the effect of feature selection with the proposed methodology are provided in Table 2. Clearly, as the feature selection set increases, the pixel-wise accuracy and Kappa score increases. It implies pixel-wise accuracy is directly proportional to our feature selection. We choose the Kappa coefficient as a performance metric because it represents the extent to which classes on the ground are correct representations of the classes on the map.

Finally, qualitative results are illustrated in Fig. 2. The ground truth tree cover of LUMS study region is 23 acres while the predicted area using the proposed model is 25 acres. It is important to note that our proposed model operating on low resolution imagery with suitable feature selection and masking operations performs better than a DeepLabv3 [7] deep learning architecture trained on high resolution imagery despite the computational complexity of the former being extremely low compared to the latter.

4. CONCLUSION

The paper proposes a methodology for estimating urban tree cover using RF classification with appropriately selected multispectral features and masking. The proposed methodology exhibits superior performance compared to classical RF classifiers that solely utilize spectral bands, as well as surpassing state-of-the-art models such as the European Space Agency (ESA) WorldCover 10m 2020 product [6] and DeepLabv3 [7] deep learning architecture trained on high resolution imagery. Our future work aims to apply the proposed technique to estimate tree cover across entire cities in Pakistan.

5. REFERENCES

[1] Titta Majasalmi and Miina Rautiainen, "Representation of tree cover in global land cover products: Finland as a case study area," *Environmental Monitoring and Assessment*, vol. 193, no. 3, pp. 1–19, 2021.

[2] Ewa Grabska, Patrick Hostert, Dirk Pflugmacher, and Katarzyna Ostapowicz, "Forest stand species mapping using the Sentinel-2 time series," *Remote Sensing*, vol. 11, no. 10, pp. 1197, 2019.

[3] Sandro Martinis, André Twele, and Stefan Voigt, "Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution terrasar-x data," *Natural Hazards and Earth System Sciences*, vol. 9, no. 2, pp. 303–314, 2009.

[4] Matthew C Hansen, Peter V Potapov, et al., "High-resolution global maps of 21st-century forest cover change," *science*, vol. 342, no. 6160, pp. 850–853, 2013.

[5] Milkessa Dangia Negassa, Demissie Tsega Mallie, and Dessalegn Obsi Gemeda, "Forest cover change detection using geographic information systems and remote sensing techniques: a spatio-temporal study on komto protected forest priority area, east wollega zone, ethiopia," *Environmental Systems Research*, vol. 9, no. 1, pp. 1–14, 2020.

[6] Ruben Van De Kerchove, Daniele Zanaga, et al., "ESA worldcover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data.," in *AGU Fall Meeting Abstracts*, 2021, vol. 2021, pp. GC45I–0915.

[7] Liang-Chieh Chen, George Papandreou, et al., "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[8] JW Rouse, RH Haas, JA Schell, and DW Deering, "Monitoring vegetation systems in the great plains with erts proceeding," in *Third Earth Reserves Technology Satellite Symposium, Greenbelt: NASA SP-351*, 1974, vol. 30103017.

[9] Alfredo Huete, Kamel Didan, et al., "Overview of the radiometric and biophysical performance of the modis vegetation indices," *Remote sensing of environment*, vol. 83, no. 1-2, pp. 195–213, 2002.

[10] "Comparasion of NDBI and NDVI as Indicators of Surface Urban Heat Island Effect in Landsat 8 Imagery: A Case Study of Iasi in: Present Environment and Sustainable Development Volume 11 Issue 2 (2017)," .

[11] Stuart K McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[12] Eva Boegh, Henrik Soegaard, et al., "Airborne multispectral data for quantifying leaf area index, nitrogen concentration, and photosynthetic efficiency in agriculture," *Remote sensing of Environment*, vol. 81, no. 2-3, pp. 179–193, 2002.

[13] Alfredo R Huete, "A soil-adjusted vegetation index (savi)," *Remote sensing of environment*, vol. 25, no. 3, pp. 295–309, 1988.

[14] Bo-Cai Gao, "NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment*, vol. 58, no. 3, pp. 257–266, 1996.

[15] Yong Zha, Jay Gao, and Shaoxiang Ni, "Use of normalized difference built-up index in automatically mapping urban areas from tm imagery," *International journal of remote sensing*, vol. 24, no. 3, pp. 583–594, 2003.

[16] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] Ramón Díaz-Uriarte and Sara Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[18] Kellie J Archer and Ryan V Kimes, "Empirical characterization of random forest variable importance measures," *Computational statistics & data analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.