Scientific Research Publishing

Journal of Geographic Information System, 2023, 15, 53-72
https://www.scirp.org/journal/jgis
ISSN Online: 2151-1969
ISSN Print: 2151-1950

Object-Based Burned Area Mapping with Extreme Gradient Boosting Using Sentinel-2 Imagery

Dimitris Stavrakoudis*, **Ioannis Z. Gitas**

Laboratory of Forest Management and Remote Sensing, School of Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: *jstavrak@auth.gr

How to cite this paper: Stavrakoudis, D. and Gitas, I.Z. (2023) Object-Based Burned Area Mapping with Extreme Gradient Boosting Using Sentinel-2 Imagery. *Journal of Geographic Information System*, **15**, 53-72.
https://doi.org/10.4236/jgis.2023.151004

Received: January 2, 2023
Accepted: February 11, 2023
Published: February 14, 2023

Abstract

The Sentinel-2 satellites are providing an unparalleled wealth of high-resolution remotely sensed information with a short revisit cycle, which is ideal for mapping burned areas both accurately and timely. This paper proposes an automated methodology for mapping burn scars using pairs of Sentinel-2 imagery, exploiting the state-of-the-art eXtreme Gradient Boosting (XGB) machine learning framework. A large database of 64 reference wildfire perimeters in Greece from 2016 to 2019 is used to train the classifier. An empirical methodology for appropriately sampling the training patterns from this database is formulated, which guarantees the effectiveness of the approach and its computational efficiency. A difference (pre-fire minus post-fire) spectral index is used for this purpose, upon which we appropriately identify the clear and fuzzy value ranges. To reduce the data volume, a super-pixel segmentation of the images is also employed, implemented via the QuickShift algorithm. The cross-validation results showcase the effectiveness of the proposed algorithm, with the average commission and omission errors being 9% and 2%, respectively, and the average Matthews correlation coefficient (MCC) equal to 0.93.

Keywords

Operational Burned Area Mapping, Sentinel-2, Extreme Gradient Boosting (XGB), QuickShift Segmentation, Machine Learning

1. Introduction

Wildfires constitute significant environmental, socio-economic, and in many cases, political pressure in Europe and, most prominently in Mediterranean countries, introducing a high risk of direct damage to humans and structures in most of the highly populated Mediterranean countries and especially in coastal regions [1]. Most wildfires in Europe—over 85% of the total burned area—take place in its Mediterranean region, where about 65,000 fires occur every year on average, burning approximately half a million hectares of wildland and forest areas [2]. The analyses performed by the European Forest Fire Information System (EFFIS) indicate an increase in the length of the wildfire season in the last 30 years, whereas the fire regime is projected to change almost everywhere in Europe in the next decades [3]. In the last few years, the pressure has been constantly increasing in northern European countries, creating significant problems due to the (comparatively) reduced preparedness and resilience.

Timely and accurate burned area mapping is essential for quantifying the environmental impact of wildfires, compiling statistics, and designing effective short- to mid-term impact mitigation measures (e.g., prevention of soil erosion or possible impacts of the fire/heavy rainfall combination). Satellite imagery has been successfully employed for mapping burned areas for several decades, since it offers a more accurate, seasonable, and resource-efficient alternative to field surveys [4], whereas it permits various levels of automation of the mapping process, especially in view of the great advancements the field of machine learning has seen the last few years. Traditionally, moderate- to coarse-resolution satellite sensors have been used for the task, such as MODIS (Moderate Resolution Imaging Spectroradiometer) [5] and MERIS (MEdium Resolution Imaging Spectrometer) [6]. These sensors offer the advantage of daily (or sub-daily) global coverage and the possibility to identify the date of the fire through a fully automated workflow. However, their coarse spatial resolution (pixel size of 500 m or greater) provides only a rough estimate of the fire perimeter.

When a detailed mapping of the affected area is required, high-resolution satellite imagery can be used instead, sacrificing the temporal frequency of sensors such as MODIS in favor of increased spatial resolution. Landsat data (having 30 m spatial resolution) have been predominately used for this purpose [7] [8] [9], due to their rich spectral information—especially the short-wave infrared (SWIR) bands they include that are important for burned area mapping—and their free data provision policy from the United States Geological Survey (USGS) since 2008 [10]. The Sentinel-2 mission—developed and operated by the European Space Agency (ESA) as part of the Copernicus programme of the European Commission (EC)—is providing free of charge high-resolution optical imagery since 2015. Sentinel-2 data are characterized by high spatial resolution (10 - 20 m, depending on the band), rich spectral information (more or less a superset of Landsat 7 ETM+ and Landsat 8 OLI bands), and high temporal frequency (5 days), characteristics which constitute them attractive for setting up an operation burned area mapping service on a national level. Several studies have investigated the potential of the Sentinel-2 data for burned area mapping [11] [12] [13] [14] [15], although much fewer than those exploiting Landsat data.

Automated approaches for mapping burned areas using high-spatial-resolution imagery typically rely on pre- and post-fire differences of appropriately selected spectral indices, enforcing some form of thresholding [16] [17] [18]. Even though most such approaches try to optimize somehow the selection of the thresholds by expert knowledge and/or empirical rules, these threshold values are sensitive to several parameters (land cover type, ecosystem type, fire severity, etc.) and it is practically very difficult to find appropriate values that would cover most of the cases. On the other hand, studies that employ well-known supervised classification approaches (e.g., [19] [20]) are very sensitive to the selection of training patterns and are very hard to scale to larger regions (e.g., on a national level).

This study presents a novel framework for mapping burned areas using pairs of Sentinel-2 imagery, aiming at fusing the advantages of the empirical approaches that rely on different spectral indices and those of the supervised classification workflow. The proposed methodology has been designed to be applicable in a nationwide operational framework, with the classification models being trained considering a database of 64 reference fire perimeters in Greece from 2016 to 2019. For the supervised classification step, the extreme Gradient Boosting (XGB) [21] machine learning framework has been selected, which the last few years showcases very high competence in achieving high accuracy in different classification and regression tasks [22]. An empirical approach for selecting representative training patterns via difference spectral indices is also proposed, which is a crucial element and guarantees the effectiveness of the whole methodology. An object-based (super-pixel) approach is also employed, to reduce the data volume and computational requirements of the algorithm.

2. Datasets Used and Preprocessing

The proposed methodology uses Sentinel-2 MultiSpectral Instrument (MSI) data. We did not consider the bands with spatial resolution of 60 m—which are primarily useful for atmospheric correction—neither the red-edge bands (bands 5, 6, and 7), which are primarily used for calculating spectral indices sensitive to vegetation stress. Moreover, we used band B8A for the near-infrared (NIR) spectral region (and not band B08), since the latter is narrow-band and more useful for identifying burned areas. The nominal specifications [23] of the six bands that were ultimately considered in this study are reported in **Table 1**.

The selected classification approach (XGB) requires a statistically adequate number of training samples for performing efficiently. Therefore, we identified 64 different wildfire events that occurred in Greece from 2016 to 2019 of different size (but greater than 50 ha), burn severity, and ecosystem type affected. For each one of these wildfires, a pre-fire and a post-fire Sentinel-2 images were downloaded from the Copernicus Open Access Hub (SciHub, https://scihub.copernicus.eu/). The dates of the pre- and post-fire images were manually selected to be as close as possible, respectively, to the date the fire

Table 1. Sentinel-2 MSI bands used in this study and their nominal specifications.

Band	Description	Spatial Resolution (m)	Central wavelength (nm)	Bandwidth (nm)
B02	Blue	10	490	65
B03	Green	10	560	35
B04	Red	10	665	30
B8A	NIR (narrow)	20	865	20
B11	SWIR 1	20	1610	90
B12	SWIR 2	20	2190	180

started and to the date it was fully controlled and, at the same time, to be cloud-free over the fire scar. Level-2A (L2A) products were downloaded, that is, atmospherically corrected images with the values representing bottom of atmosphere (BoA) reflectance. For the 2016 products (and for a few 2017 ones) that only Level-1C products were available via SciHub, the latter were converted to L2A by running the Sen2Cor processor [24] in house. A total of 103 different tiles (Sentinel-2 images) have been downloaded and processed. This number is smaller than the double of the wildfire events, since some cases required the same pre-fire or post-fire image, or both, with other ones. For each image, the raster bands are stacked and upsampled to the finer spatial resolution (*i.e.*, 10 m), considering a nearest neighbor resampling. Moreover, they are converted to reflectance values in [0, 1] (from the original range of [0, 10,000]) and saturated in this range. Clouds are masked out (*i.e.*, marked as no-data) using the scene classification band of Sentinel-2 L2A products (we consider the cloud medium and high probability values of the band). We apply a morphological opening operator to the initial cloud mask with a circular structuring element of 10 pixels radius, followed by a morphological dilation (buffer) of 10 pixels. This process creates a more smooth and safe cloud mask, compared to the original scene classification band. Finally, non-land areas (identified using the Greece's official land area layer) are also marked as no-data regions and excluded from further processing.

The reference fire perimeters were manually delineated from the pre- and post-fire image pairs, through careful visual inspection of the images and derivative spectral indices. Although we tried to be as precise as possible, minor omissions may have occurred during this process, since the 10 m spatial resolution (for four bands) of Sentinel-2 sets a limit to the level of detail that can be observed, especially over small features/areas (most notably, agricultural fields). Nevertheless, these inconsistencies are relatively very small compared to the actual burned areas and the reference dataset can be considered generally correct. Figure 1 depicts the distribution of the reference fire perimeters across time and space. The selected wildfires are generally spread across the country, with a bias towards Central Greece, which is historically the most fire-prone region. The

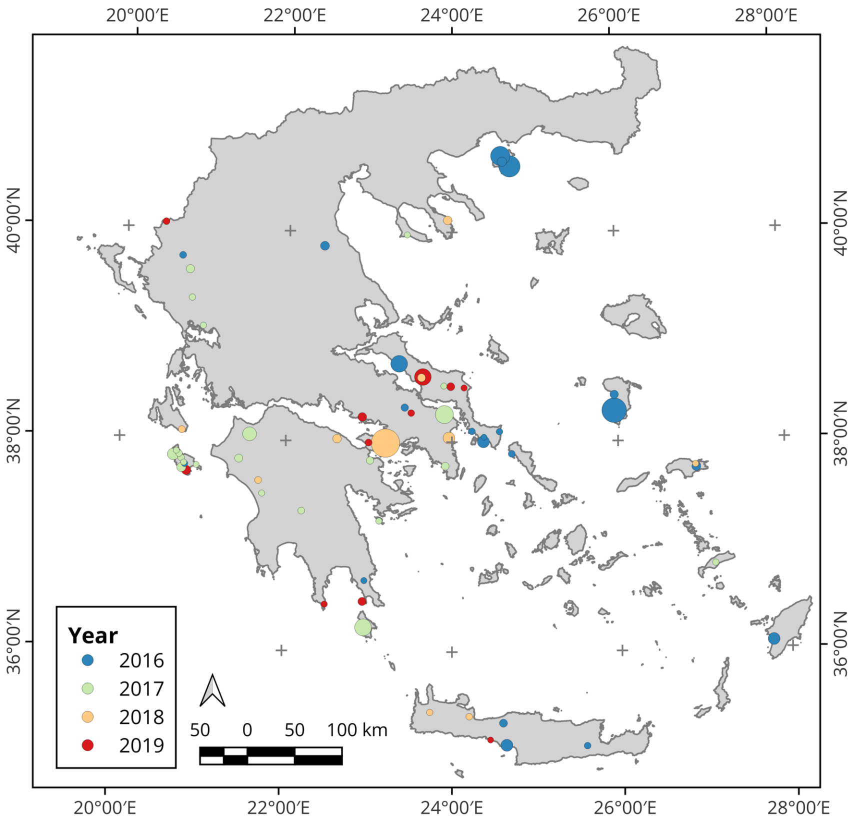


Figure 1. Location of the 64 wilfire events in Greece from 2016 to 2019 used as the reference set. The diameter of each point represents the relative burned area, with the largest corresponding to 5376.89 ha and the smallest to 71.23 ha.

size of each point in the figure represents the relative burned area, largest corresponding to 5376.89 ha and the smallest to 71.23 ha.

3. Methodology

Figure 2 depicts the workflow of the employed methodology. First, the post-fire image is used to produce a segmentation of an area around the fire scar, via the QuickShift algorithm. The objects are used to calculate object-level features (mean values) from both the pre-fire and post-fire images. Subsequently, additional spectral images at object level are calculated for both images, as well as their difference. The difference indices are used to select a subset of samples from the reference set, which formulate the training set. The latter is used to train an XGB classifier, considering as inputs features from the pre- and post-fire bands and the difference indices. The train classifier is subsequently employed to product the burned area map. The rest of this section details each step of the methodology.

3.1. Image Segmentation

To a large extent, the traditional OBIA approach [25] [26] tries to optimize the selection of the segmentation's scale, meaning to derive the largest possible

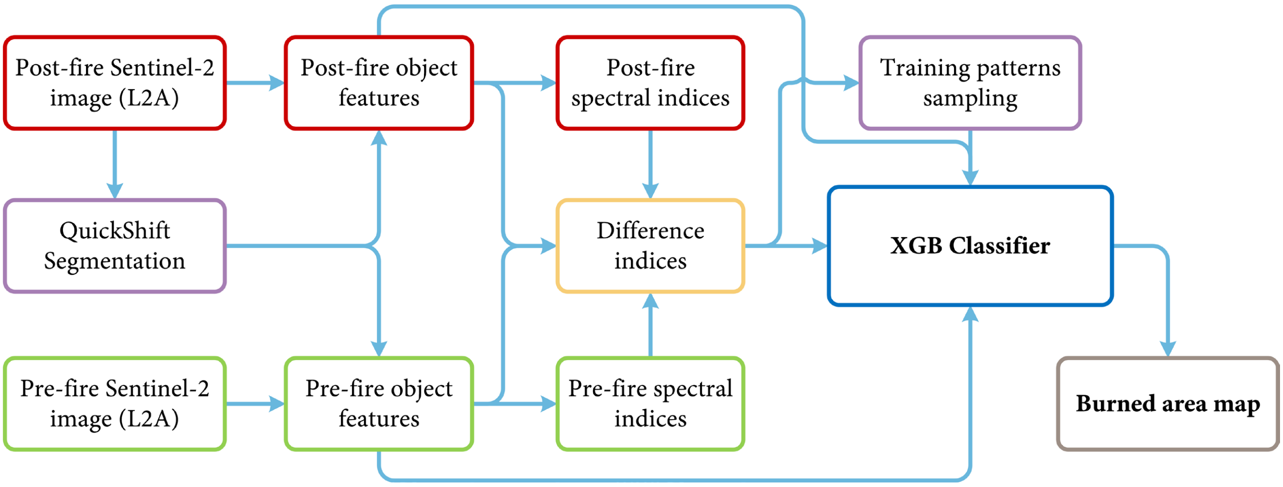

Post-fire Sentinel-2 image (L2A)
Post-fire object features
Post-fire spectral indices
Training patterns sampling
QuickShift Segmentation
Difference indices
XGB Classifier
Pre-fire Sentinel-2 image (L2A)
Pre-fire object features
Pre-fire spectral indices
Burned area map

Figure 2. Schematic workflow of the proposed methodology.

objects that are homogeneous at the same time, two inherently contradicting objectives. In our work, we follow an object-based approach simply to compress the data volume, since the training set is very large. As such, we prefer to over-segment the image (*i.e.*, create objects of a few pixels or few dozens of pixels), leaning more towards the principles of super-pixel segmentation [27] rather those of the traditional OBIA. An over-segmentation reduces the risk of misclassifications due to an incorrect local scale but significantly compresses the volume of data to be processed.

For segmenting the image, we use the QuickShift segmentation algorithm [28], which is a significantly faster variant of the Mean-Shift segmentation algorithm [29]. Although QuickShift can be employed to multidimensional datasets, using several features is both computationally inefficient and practically unnecessary, whereas it also increases the algorithm's sensitivity. Hence, we use only the visible (B02, B03, B04) and NIR (B8A) bands of the post-fire image for the segmentation process. The pre-fire image is deliberately not used for this process, since it could lead to erroneous segmentations, given that the fire scar exists only in the post-fire image. Moreover, the Sentinel-2 tiles are first cropped in a region around each reference fire perimeter. Specifically, we estimate the bounding box of each fire perimeter and we double this box in each axis, extending by half in each direction. Hence, the region of analysis is four times larger (unless confined by the original Sentinel-2 tile's bounds) than the fire perimeter's bounding box (double in each dimension) and centered at the latter's center. In an operational workflow, we assume that the user supplies this extended region of analysis, in an area around the fire scar. The Conclusions section discusses this issue further, in terms of applicability.

The scale of the segmentation in QuickShift is controlled mostly by three parameters: ratio, kernel size, cutoff point for data distances. Based on preliminary experiments with a few images, the values of these parameters have been selected to lead to over-segmentation (the selected values were 5, 5, and 2, respectively).

But for this process to be consistent, all images must be in a similar range of values. To enforce this, we first saturate the bands in the range [0, 0.4] and then linearly scale this range to [0, 255] and save the bands in 8-bit unsigned integer pixel type. Note that the selected range will only saturate non-burned bright areas, since the burned areas have very low reflectance values in the four bands. This is also the range commonly employed for visualizing the images.

For each object, the average value of the pixels belonging to it in each band is used as the object-level features for the subsequent analysis. Moreover, a truth value for each object (burned or unburned) is derived from the reference fire perimeters. The fire perimeters are first rasterized in the cropped images' resolution and the object-level truth value is determined as the most frequent value among the object's pixels. The segmentation process results in a total of over 2 million objects (2,074,732 to be exact), with 1,786,921 labeled as unburned and 287,811 as burned. The number of unburned objects is much larger than that of the burned, because the fire scar is also typically smaller than its perimeter's bounding box (since its shape is irregular) and the extended region of analysis around each perimeter is four times this bounding box (unless for cases close the original image's bounds). Even if the final dataset seems large, it is still smaller than the original one, approximately 18 times smaller (the number of pixels in all objects is close to 38 million).

3.2. Spectral Indices

Four spectral indices frequently used in burned area mapping studies have been calculated, which are reported in Table 2. Most of them employ the classical normalized difference formulation of the Normalized Difference Vegetation Index (NDVI) [30], but using one or both of the SWIR bands, which are sensitive to moisture content and consequently facilitate the discrimination of the burned areas. The indices are calculated in both the pre- and post-fire cropped images and at the object level, *i.e.*, using the mean object value of the respective band. For each spectral index *SI*, the difference between the pre-fire and post-fire values is also calculated. The difference, which is denoted as d*SI* for any index *SI* (e.g., dNBR when using NBR as *SI*), is calculated as pre-fire minus post-fire value, *i.e.*, $\mathrm{d}SI = SI_{\mathrm{pre}} - SI_{\mathrm{post}}$.

Table 2. Spectral indices considered in this study. The Sentinel-2 band names are those reported in Table 1.

Acronym	Name	Equation	Introduced in
NBR	Normalized burn ratio	$(\mathrm{B8A} - \mathrm{B12})/(\mathrm{B8A} + \mathrm{B12})$	[31]
NBR2	Normalized burn ratio 2	$(\mathrm{B11} - \mathrm{B12})/(\mathrm{B11} + \mathrm{B12})$	[32]
MIRBI	Mid-infrared burn index	$10 \cdot \mathrm{B12} - 9.8 \cdot \mathrm{B11} + 2$	[33]
NDII	Normalized difference infrared index	$(\mathrm{B8A} - \mathrm{B11})/(\mathrm{B8A} + \mathrm{B11})$	[34]

3.3. Selection of Training Samples

The sample selection process is the most crucial part of our methodology. The objective is to avoid using an excessive number of training patterns that would be computationally inefficient, but we need to include representative cases in the training set. Randomly sampling patterns from a comparatively much larger pool cannot guarantee the latter, which leads to extreme overfitting and, hence, poor accuracy.

The difference spectral indices are commonly used for discriminating burned scars in a simple manner, with the most well-known example being the ubiquitous use of dNBR for this purpose, which has also been shown to be highly correlated with burn severity [31]. As explained in the introductory section, the problem of such approaches is that the threshold value for discriminating burned from unburned areas varies significantly in different scenarios (hence, it requires manual determination by the user) and that several other land use changes can produce a dNBR (or any other index) value like that of burned areas. Yet, the distribution per class (burned or unburned) of the difference indices can serve as an efficient way to sample training patterns and then train a machine learning classifier, to overcome these limitations.

Following this reasoning, our methodology selects training patterns from the distribution of a difference index d*SI*. Obviously, we need to include "clear" objects in the training set, meaning objects that are easily discriminable as unburned or burned (but considering all different land cover scenarios observed in practice). However, it is also important to include the ambiguous cases, *i.e.*, objects that have similar d*SI* values but belong to different classes, as well as patterns of the one class with values in ranges typically dominated by the other class. These ambiguous patterns are the ones missed by purely random approaches and produce the misclassifications of the simple approaches of thresholding a difference index.

Our approach starts considering a difference index d*SI* and a pool of provisional training objects. For each of the two classes, the histogram of the provisional training objects represents the distribution of the patterns in the space of d*SI*. The left class is defined as the one with the lower values in general, *i.e.*, as the one with the lower median value, whereas the other is referred as right class. For most (although not for all) band or spectral index differences considered here, the unburned class is the left one. The two distributions are sampled independently. Without, loss of generality, we will describe the process for sampling the left distribution. A substantial quantile value for the left distribution (q_ℓ) determines the threshold after which (*i.e.*, on the right) the left class is intermixed with the clear areas of the right distribution. We selected the 0.99-quantile for this purpose (corresponding to a value t_ℓ on the space of d*SI*), meaning that 1% of the left distribution's samples have d*SI* values larger than t_ℓ. For the right class, the complementary quantile ($q_r = 1 - q_\ell = q_{0.01}$, corresponding to a value t_r of d*SI*) represents the opposite, *i.e.*, 1% of the right distribution's samples have d*SI* values less than t_r. d*SI* values less than t_ℓ and greater than t_r are considered to belong in the clear regions of the left ($\mathcal{C}_\ell$) and right ($\mathcal{C}_r$) classes, respectively. The region $\mathcal{A} \equiv [t_r, t_\ell]$ is considered the ambiguous region of the difference index. If $t_\ell < t_r$, then the ambiguous area is empty. Yet, this never occurs in practice, because if it did, then the two classes could be discriminated without any errors. Supposing we are sampling *N* training samples from the left distribution, $p_{\mathcal{A}} \cdot N$ will be sampled randomly from the ambiguous area $\mathcal{A}$. In the experiments we set $p_{\mathcal{A}} = 0.1$, *i.e.*, 10% of the samples are drawn from the ambiguous area and the remaining 90% from the clear region $\mathcal{C}_\ell$ (ratio $p_{\mathcal{C}} = 1 - p_{\mathcal{A}}$). The latter (which is the region with d*SI* values in $[t_{\min}, t_\ell)$, with $t_{\min}$ being the minimum value of d*SI*), is further split into *k* quantiles (selected $k = 10$ in the experiments) and an equal number of patterns ($p_{\mathcal{C}} \cdot N/k$) is randomly sampled from each quantile range. In each such as subregion, an additional proportion p_m (set as $p_m = 0.1$ in the experiments) of samples from the right distribution—with respect to those sampled from the left distribution in that range, *i.e.*, $p_m \cdot p_{\mathcal{C}} \cdot N/k$—is also randomly sampled from the subregion, bounded from the top to the number of samples of the right distribution (*i.e.*, if there are patterns available to sample). The latter are considered because they are the patterns of the competing class (*i.e.*, the right class) but within the left distribution's clear area and should therefore be taken into consideration by the classifier, to account for the classes' intermixing. They are also added to the patterns sampled from the left distribution, so the actual number of patterns sampled is greater than *N*. The theoretical maximum number of patterns sampled is $(1 + p_m \cdot p_{\mathcal{C}}) \cdot N$, although practically it is closer to *N*, since there are usually not that many patterns from the right class in all quantile ranges of the left distribution's clear area. With the numbers for the parameters selected for the experiments, the above means that 10% of left class samples are sampled from the ambiguous area, another 90% of left class samples from the clear area (9% from each region) and an additional $0.09 \cdot N$ of right class patterns (at maximum). The process is repeated for the right distribution sampling slightly more than *N* additional patterns. The steps are the same, inverting all terms (*i.e.*, equivalent to applying the previous process in the negated values of the right class patterns). For the experimentation setup, we set *N* = 25,000, which means that slightly more than 50,000 patterns have been sampled to formulate the training set, equally distributed between the two classes.

The process above ensures that the training patterns are sampled from all parts of the burned and unburned objects' distributions, thus representing all differences observed in the two images. Moreover, special care is taken for the ambiguous cases, *i.e.*, differences between the two images that can belong to either class. Although the previous description may seem complicated, the process is very easy to implement and can be easily explained with a figure. **Figure 3(a)** shows the distribution of the unburned class's object values of the dNBR index. The distribution is shown as the filled area and as a normalized (inverse) cumulative

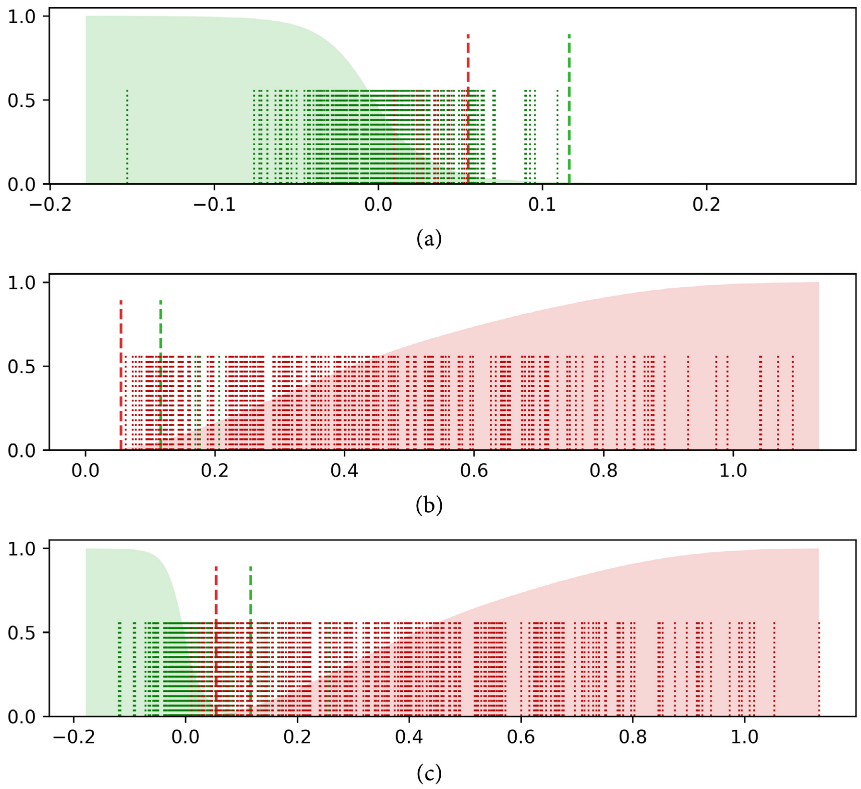


Figure 3. Example of sample selection process, sampling 200 patterns from each class and considering the dNBR index. Green colors represent the unburned class, whereas red colors the burned one. (a) Cumulative histogram of the unburned (left) distribution, mixed area thresholds and patterns sampled, (b) the respective plot for the burned (right) distribution, and (c) the respective plot for both distributions.

histogram. That is, the values on the horizontal axis are dNBR values, whereas the values on the vertical axis are the cumulative proportion of patterns, integrating from right to left (*i.e.*, an inverse density distribution). In dNBR, unburned objects have a median value much lower than burned ones and, hence, the unburned distribution is towards the left side (decreasing dNBR values). The red and green dashed lines are, respectively, the t_r and t_ℓ. Hence, 1% of the unburned objects have dNBR values greater than the green dashed line and 1% of the burned objects have dNBR values less than the red dashed line. The ambiguous area ($\mathcal{A}$) is the area between these two dashed lines. For clarity in the presentation, in this example we set $N = 200$ samples to be drawn from each distribution. A proportion of $p_{\mathcal{A}} = 0.1$ of patterns (20 in this case) are sampled from this area and are shown with the shorter dotted line (each dotted line is an object drawn to be included in the final training set). The remaining are sampled from the clear region of unburned distribution (left of the red dashed line), but in regions determined by 10 quantiles. Hence, comparatively fewer samples are drawn from the far left, because the proportion of objects with smaller values is much smaller; the cumulative distribution is almost flat for dNBR values less than approximately -0.1, meaning that almost all objects have greater values.

Some additional patterns are sampled from the burned class as well (the red dotted lines in between), due to the p_m proportion explained above. Figure 3(b) shows the sampling of the burned distribution, out of which an additional $N = 200$ objects are sampled (this time, belonging to the burned class), plus the few additional unburned samples in between. Note that the ambiguous region is the same in the two plots, the optical difference arises from the fact that the burned class has a distribution with much longer tails (greater variability in dNBR values) than the unburned class. Figure 3(c) combines the two previous plots, showcasing how both the burned and the unburned classes are sampled appropriately in their whole range of values.

All individual band differences (dB02, dB03, etc.) and all differences of the spectral indices in Table 2 (dNBR, dNBR2, dMIRBI, and dNDII) are considered as candidates for sampling the training patterns. However, the feature with the highest separability in discriminating the two classes is expected to produce better results. If the distributions of the two classes in a feature have high degree of overlap, that means that the differences in the two images are small, despite the area having been burned. Such a feature is not suitable for appropriately sampling training patterns, since the differences in land cover are not visible anyway and the sampling will be close to purely random. An easy way to estimate a separability measure is the proportion of objects in each distribution that belong to its clear region. With reference to Figure 3(c), this is the region left of the red dashed line for the unburned distribution (green, left) and the region right of the green dashed line for the burned distribution (red, right). The separability is defined in this work as the sum of the objects' proportions in those two regions. Formally, for any difference feature dSI, let M_ℓ be the total number of samples of the left distribution and M_C^ℓ the number of samples with value less than t_r (the dSI value corresponding to the 0.01-quantile of the right distribution). Also, M_r is the total number of samples of the right distribution and M_C^r the number of samples with value greater than t_ℓ (the dSI value corresponding to the 0.99-quantile of the left distribution). Then, the separability index of feature dSI is defined as:

$$\operatorname{sep}(\mathrm{d}SI) = \frac{M_C^\ell}{M_\ell} + \frac{M_C^r}{M_r}. \tag{1}$$

With this definition, a separability value of 0 means that the two classes are fully discriminated (no ambiguous areas) in this feature and a value of 2 means that the histograms of the two distributions are identical (no difference between values of the dSI between the classes). Ultimately, the difference index with the lowest separability index value is used for selecting the samples of the training set, according to the sampling processes described previously. In all folds performed in the experiments, dNBR was always selected as the best index, followed by the other difference spectral indices. Band differences had even higher index values and especially the bands in the visible portion of the spectrum (B02, B03, and B04) had index values close to 2.

3.4. Training and Applying the XGB Classifier

Having selected the training samples, training the classifier is rather straightforward. The XGB framework been selected because of its showcased competence in achieving high accuracy in different classification and regression tasks [22], including very recently in wildfire research [35] [36] [37] (in quite different tasks than our work). Perhaps most importantly, the XGBoost open-source software library [21] is extremely optimized in terms of parallel execution and memory minimization, being able to train an extremely deep set of trees using more than 60,000 training patterns (as we selected for the experiments) in a few seconds. As classification features, the (object-level) values for the pre- and post-image Sentinel-2 bands reported in **Table 1** have been used, in addition to all difference spectral indices (dNBR, dNBR2, dMIRBI, dNDII).

The XGBoost algorithm has several hyper-parameters that control overfitting and ultimately its ability to classify correctly unseen data. A space of possible values was selected for the most important of those (7 in total hyper-parameters) and the optimal combination was found via a randomized cross-validation (CV) search. For this purpose, the Python package scikit-optimize [38] was used. The latter employs a Bayesian optimization framework for optimizing the randomized CV search, to avoid the computationally demanding exhaustive search.

After training, the classification model is applied to the testing objects (*i.e.*, unseen during the training process) and the final thematic map is produced rasterizing the objects' predicted value to the original image resolution (within the region of analysis for each fire perimeter).

3.5. Experimental Setup

The experimental workflow followed a 5-fold cross-validation (CV) procedure. If, however, we used the full pool of segmented objects (the 2,074,732 objects mentioned in Section 3.2) for the CV, we would derive biased results (*i.e.*, inflated accuracy metrics), because objects from the same wildfire event would be used for both training and testing. Because of the nature of our problem, all objects from a single fire event should belong to either the training set or the testing one in each CV repetition. Moreover, the burned area of the reference perimeters should be taken into consideration, since the there is significant variably in the area burned between the 64 available fire perimeters.

To respect these constraints, we employed a group 5-fold CV procedure, taking the reference burned area also into consideration. More specifically, the list of the 64 reference perimeters were randomly shuffled and partitioned into 5 non-overlapping parts (folds), so that the sum of the reference burned areas was approximately equal in the five folds. In each one of the 5 repetitions of the 5-fold CV procedure, all objects belonging to the respective fold were used as the testing dataset, whereas all other objects serve as the (provisional) training dataset. And the process is repeated in a circular fashion, so that all folds formulate the testing dataset at some repetition. Though this approach, each of the 64 reference perimeters considered belongs—as a whole—to a testing fold at one of the 5 iterations.

For each training (4 folds) and testing (the remaining fold) pairs, the objects in the training set are considered as provisional training samples and the whole process described in Sections 3.3 and 3.4 is applied. More specifically, the separability index in Equation (1) was calculated for all difference indices, the feature with the lowest value was selected, and then used to sample the classifier's training set via the procedure explained in Section 3.3. We sampled $N = 25,000$ patterns from each distribution, therefore the classifier's training set comprised slightly more than 50,000 training points (for the reasons explained in Section 3.3). In all cases, the pool of available provisional training objects (all objects in the 4 folds) was much larger than this number for both classes. Subsequently, the XGB classifier was trained for each repetition of the 5-fold CV anew, meaning that its hyper-parameters were also determined a new via the randomized CV search procedure described in Section 3.4. Hence, the classification models in the 5 repetitions are different. Each trained model was finally employed to classify the testing objects, resulting in the testing (predicted) burned areas maps. Merging the testing folds' results from the 5 repetitions we can also recreate (if desired) the whole initial database of 64 wildfire events as a testing one, although the mappings have been produced by different XGB models.

The whole process was implemented in the Python programming language and executed on a laptop personal computer with an Intel® Core™ i5-12500H processor at 2.50 GHz and 16 GB of RAM. Even with this rather medium-capability configuration and with over 50,000 training samples, finding the hyper-parameters and training each XGB model required less than one minute, whereas applying the classifier to obtain the prediction required less than one second.

For quantitative analysis, measures derived from the confusion matrix were calculated, namely, overall accuracy (O. A.), precision, recall, F_1 score, and Matthews correlation coefficient (MCC) [39], as well as the area under the curve (AUC) score from the receiver operating characteristic (ROC) analysis [40]. In calculating these metrics, the positive (P) class was considered the burned class, which is the class of interest. Because the dataset is imbalanced, metrics such as the O. A. and AUC are expected to appear inflated, since the number of unburned samples is much higher, and they are rarely misclassified as burned. Therefore, the precision, F_1 and MCC metrics are more suitable for assessing the methodology's effectiveness. The precision is linked to commission errors, *i.e.*, the unburned objects misclassified as burned, which is of high importance in burned area mapping studies. All metrics are normalized in [0, 1], with the values of towards 1 meaning higher accuracy.

4. Results

Table 3 reports the values of the accuracy metrics for the 5-folds, along with the average and standard deviation values, when the objects in the testing fold are

Table 3. Accuracy metrics for the 5-fold CV, when considering each object in the testing fold as a individual patterns.

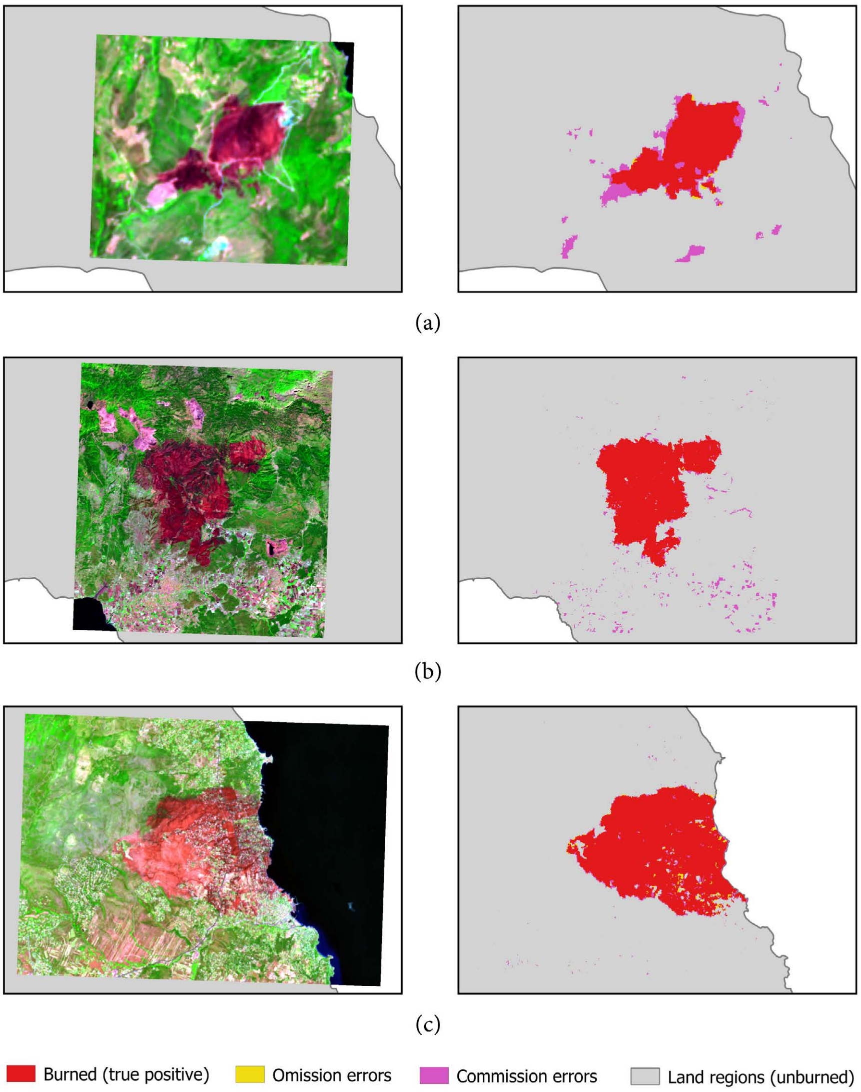
Fold	O. A.	Precision	Recall	F_1	MCC	AUC
1	0.99	0.94	0.99	0.97	0.96	0.99
2	0.99	0.95	0.99	0.97	0.96	0.99
3	0.98	0.94	0.95	0.95	0.94	0.97
4	0.99	0.90	0.99	0.94	0.94	0.99
5	0.99	0.92	0.99	0.95	0.95	0.99
Average	**0.99**	**0.93**	**0.98**	**0.96**	**0.95**	**0.99**
Std	*0.003*	*0.019*	*0.016*	*0.011*	*0.012*	*0.008*

O. A.: overall accuracy; Std: standard deviation.

considered as individual patterns. The latter is the typical way of assessing the classification accuracy of a classifier. In our case, however, each such individual pattern is an object in the image and objects have difference size (area). This is not accounted in this analysis (but treated below in this section). Nevertheless, these results are representative of the classifier's performance, since the classifier considers the input object features as individual points (patterns). The results prove XGB's efficiency in producing highly accurate models, with the average precision being 93% and MCC and F_1 scores greater or equal to 95%. Therefore, the sampling scheme of Section 3.3 proves to be efficient. Moreover, it is evident that the XGB framework can effectively exploit the large volume of information it is provided with.

As mentioned previously, each object has a difference size in general. The results presented so far do not account for that. To this end, we produced the burned area map from each testing partition in the 5-fold CV and merge these testing maps. To do so, all objects belonging to a given fire event were used do derive a raster map, by assigning to all pixels belonging in an object the predicted value of the classifier for this object. Then, each such burned area map is compared with its (rasterized) reference perimeter and the accuracy metrics are recalculated. Table 4 reports the results of this analysis. Because it would require 64 rows for the analysis, we report instead the aggregate measures, *i.e.*, average, standard deviation, and minimum and maximum values observed within those 64 maps. The results are generally consistent with the previous ones, with the precision, F_1 and MCC metrics having slightly lower values. The latter is observed because a few cases with lower accuracies (e.g., a single fire mapping exhibits precision of 76% but most of the other maps have more than 85%). Most frequently, the cases with reduced accuracies are the smaller wildfire events, since a smaller burn scar increases the sensitivity of the metric (a few small misclassifications can affect the metrics to a greater extent than in large wildfires).

To give a visual assessment of the results, Figure 4 presents three examples of testing burned area maps produced via the proposed methodology. The left


(a)

(b)

(c)

Figure 4. Three examples of the burned mapping results. The maps to the left show the (cropped) post-fire Sentinel-2 images, using a false-color composite with B12, B8A and B04 in lieu of RGB, respectively. The maps on the right show the same region with the left column (in each row of plots), but with the burned maps overlaid on shapefile with the official Greece's land areas. To save space, the omission (yellow) and commission (magenta) errors are also shown along with the area mapped as burned (red). (a) The case exhibiting the lowest precision score, (b) a large wildfire with below slightly below average precision score, and (c) a large wildfire with high precision score.

Table 4. Accuracy metrics for the 5-fold CV, when considering the burned areas produced in comparison to the reference perimeters.

	O. A.	Precision	Recall	F_1	MCC	AUC
Average	0.98	0.91	0.98	0.94	0.93	0.98
Std	*0.009*	*0.049*	*0.016*	*0.029*	*0.031*	*0.009*
Minimum	0.96	0.76	0.91	0.86	0.84	0.95
Maximum	1.00	0.97	1.00	0.98	0.97	0.99

O. A.: overall accuracy; Std: standard deviation.

column of maps shows the (cropped) post-fire Sentinel-2 images in each case, using a false-color composite with B12, B8A and B04 in lieu of RGB, respectively, traditionally used for easily visualizing burned areas. The right column of maps shows the classification result, simply over Greece's land area (in each row, the extent of two maps is identical). To save space and make the maps clearer, the reference perimeter is not shown. Instead, the maps on the right show the pixels labeled as burned by the classifier (red color), the omission errors (pixels labeled as unburned but denoted as burned by the reference set) with a yellow color, and the commission errors (pixels incorrectly labeled as burned) with a magenta color. With this view, the pixels labeled as unburned by the classifier do not need to be shown (they are all others in the cropped region of analysis).

Figure 4(a) depicts the result exhibiting the lowest precision score (76%). It is the wildfire with the smallest area in the reference set (only 71.23 ha). The comparatively low precision score is due to unburned areas misclassified as burned, mostly in agricultural fields, with the largest area touching the fire perimeter. The inflated (compared to other cases) precision error is a result of the small area of the burned area (since it is a proportion). Figure 4(b) depicts the result for a large wildfire (2614.19 ha) exhibiting a precision score close but still lower than the average, *i.e.*, 88%. The omission errors are negligible, in very small areas within the fire perimeter. The lower-than-average precision score is due to commission errors in nearby agricultural fields, which a problem encountered in all automated burned area algorithms. Note that most of these misclassifications could be eliminated through a simple filtering process. For example, it is unrealistic to expect that a fully automated algorithm that exploits the 10 - 20 m spatial resolution of Sentinel-2 could provide reliable results for burned areas less than 1 ha or even 5 ha. Therefore, connected components less than a similar threshold could be automatically eliminated, greatly reducing the errors. Such small patches are usually not useful for the stakeholders interested in burned area mapping products, since they are typically forest management bodies that are concerned with the environmental and socio-economic impacts of much larger burned patches. Nevertheless, such a filtering process is not considered in this work since it is outside of its objectives. Finally, Figure 4(c) depicts the result for a relatively large wildfire (1442.47 ha) of 2018 exhibiting a high precision score of 95%. This was the most devastating wildfire in Greece (and Europe) in terms of deaths toll. Despite being in a highly complex intermix wildland urban interface (WUI) region, the classifier exhibits negligible misclassifications.

The above visual examples showcase the proposed methodology's effectiveness in automatically deriving a burned area map, but also highlight its limitations and problems. Similar observations were made for the rest of the wildfire mappings, with the misclassifications being mostly due to agricultural fields or barren areas misclassified as burned.

5. Conclusions and Future Work

The herein proposed methodology combines the efficiency of the XGB machine learning framework with a simple yet efficient empirical process for guiding the burned area mapping process and ensuring its effectiveness. The classification workflow is automated, but it still requires user intervention, in selecting the pre- and post-fire Sentinel-2 images and manually defining a region of analysis around the burn scar. By itself, this process is useful in many operational workflows. The manual work can be reduced by exploiting online downstream services available today (e.g., Sentinel Hub), reducing these manual processes to a matter of a few minutes.

Nevertheless, the methodology presented here can also serve as a basis for fully automating the process, with only leaving to the user the responsibility to check the results. For example, with each new Sentinel-2 image becoming available, a trained classifier could be employed to constantly map potential burned areas. A pool of previous images of the same Sentinel-2 tile would be needed to be stored for this purpose (to serve as the pre-fire image) or the most recent cloud-free composite image from previous dates could be possibly continuously produced and kept. In either case, a very efficient cloud masking algorithm would be required. The cloud masking process employed in this work is a simple one and, although it removes most of the clouds, it is not perfect and does not treat cloud shadows at all. However, some highly efficient cloud-masking algorithms have recently become available for Sentinel-2 imagery, such as s2cloudless. In addition, some context-based issues would need to be resolved, such as how to merge subsequent mapping of the same wildfire, which were split because the fire scar was partially obscured by clouds on the first date. We intend to focus on resolving these issues in our future work, to turn the herein proposed methodology into a fully automated server-side procedure.

Acknowledgements

This research is co-financed by Greece and the European Union (European Social Fund—ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Reinforcement of Postdoctoral Researchers—2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Pausas, J.G., Llovet, J., Rodrigo, A. and Vallejo, R. (2009) Are Wildfires a Disaster in the Mediterranean Basin?—A Review. *International Journal of Wildland Fire*, **17**, 713-723. https://doi.org/10.1071/WF07151

[2] San-Miguel-Ayanz, J., Moreno, J.M. and Camia, A. (2013) Analysis of Large Fires in European Mediterranean Landscapes: Lessons Learned and Perspectives. *Forest Ecology and Management*, **294**, 11-22. https://doi.org/10.1016/j.foreco.2012.10.050

[3] San-Miguel-Ayanz, J., Durrant, T., Boca, R., Libertà, G., Branco, A., de Rigo, D., *et al.* (2018) Forest Fires in Europe, Middle East and North Africa 2017. Publications Office of the European Union, Luxembourg.

[4] Chuvieco, E., Mouillot, F., van der Werf, G.R., San Miguel, J., Tanase, M., Koutsias, N., *et al.* (2019) Historical Background and Current Developments for Mapping Burned Area from Satellite Earth Observation. *Remote Sensing of Environment*, **225**, 45-64. https://doi.org/10.1016/j.rse.2019.02.013

[5] Giglio, L., Boschetti, L., Roy, D.P., Humber, M.L. and Justice, C.O. (2018) The Collection 6 MODIS Burned Area Mapping Algorithm and Product. *Remote Sensing of Environment*, **217**, 72-85. https://doi.org/10.1016/j.rse.2018.08.005

[6] Chuvieco, E., Yue, C., Heil, A., Mouillot, F., Alonso-Canas, I., Padilla, M., *et al.* (2016) A New Global Burned Area Product for Climate Assessment of Fire Impacts. *Global Ecology and Biogeography*, **25**, 619-629. https://doi.org/10.1111/geb.12440

[7] Goodwin, N.R. and Collett, L.J. (2014) Development of an Automated Method for Mapping Fire History Captured in Landsat TM and ETM + Time Series across Queensland, Australia. *Remote Sensing of Environment*, **148**, 206-221. https://doi.org/10.1016/j.rse.2014.03.021

[8] Bastarrika, A., Alvarado, M., Artano, K., Martinez, M.P., Mesanza, A., Torre, L., *et al.* (2014) BAMS: A Tool for Supervised Burned Area Mapping Using Landsat Data. *Remote Sensing*, **6**, 12360-12380. https://doi.org/10.3390/rs61212360

[9] Hawbaker, T.J., Vanderhoof, M.K., Beal, Y.-J., Takacs, J.D., Schmidt, G.L., Falgout, J.T., *et al.* (2017) Mapping Burned Areas Using Dense Time-Series of Landsat Data. *Remote Sensing of Environment*, **198**, 504-522. https://doi.org/10.1016/j.rse.2017.06.027

[10] Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., *et al.* (2008) Free Access to Landsat Imagery. *Science*, **320**, 1011. https://doi.org/10.1126/science.320.5879.1011a

[11] Huang, H., Roy, D.P., Boschetti, L., Zhang, H.K., Yan, L., Kumar, S.S., *et al.* (2016) Separability Analysis of Sentinel-2A Multi-Spectral Instrument (MSI) Data for Burned Area Discrimination. *Remote Sensing*, **8**, Article No. 873. https://doi.org/10.3390/rs8100873

[12] Fernández-Manso, A., Fernández-Manso, O. and Quintano, C. (2016) SENTINEL-2A Red-Edge Spectral Indices Suitability for Discriminating Burn Severity. *International Journal of Applied Earth Observation and Geoinformation*, **50**, 170-175. https://doi.org/10.1016/j.jag.2016.03.005

[13] Amos, C., Petropoulos, G.P. and Ferentinos, K.P. (2019) Determining the Use of Sentinel-2A MSI for Wildfire Burning & Severity Detection. *International Journal of Remote Sensing*, **40**, 905-930. https://doi.org/10.1080/01431161.2018.1519284

[14] Roteta, E., Bastarrika, A., Padilla, M., Storm, T. and Chuvieco, E. (2019) Development of a Sentinel-2 Burned Area Algorithm: Generation of a Small Fire Database for Sub-Saharan Africa. *Remote Sensing of Environment*, **222**, 1-17. https://doi.org/10.1016/j.rse.2018.12.011

[15] Roy, D.P., Huang, H., Boschetti, L., Giglio, L., Yan, L., Zhang, H.H., *et al.* (2019) Landsat-8 and Sentinel-2 Burned Area Mapping—A Combined Sensor Multi-Temporal Change Detection Approach. *Remote Sensing of Environment*, **231**, Article ID: 111254. https://doi.org/10.1016/j.rse.2019.111254

[16] Pulvirenti, L., Squicciarino, G., Fiori, E., Fiorucci, P., Ferraris, L., Negro, D., *et al.* (2020) An Automatic Processing Chain for Near Real-Time Mapping of Burned

Forest Areas Using Sentinel-2 Data. *Remote Sensing*, **12**, Article No. 674. https://doi.org/10.3390/rs12040674

[17] Woźniak, E. and Aleksandrowicz, S. (2019) Self-Adjusting Thresholding for Burnt Area Detection Based on Optical Images. *Remote Sensing*, **11**, Article No. 2669. https://doi.org/10.3390/rs11222669

[18] Sertel, E. and Alganci, U. (2016) Comparison of Pixel and Object-Based Classification for Burned Area Mapping Using SPOT-6 Images. *Geomatics*, *Natural Hazards and Risk*, **7**, 1198-1206. https://doi.org/10.1080/19475705.2015.1050608

[19] Colson, D., Petropoulos, G.P. and Ferentinos, K.P. (2018) Exploring the Potential of Sentinels-1 & 2 of the Copernicus Mission in Support of Rapid and Cost-Effective Wildfire Assessment. *International Journal of Applied Earth Observation and Geoinformation*, **73**, 262-276. https://doi.org/10.1016/j.jag.2018.06.011

[20] Brown, A.R., Petropoulos, G.P. and Ferentinos, K.P. (2018) Appraisal of the Sentinel-1 & 2 Use in a Large-Scale Wildfire Assessment: A Case Study from Portugal's Fires of 2017. *Applied Geography*, **100**, 78-89. https://doi.org/10.1016/j.apgeog.2018.10.004

[21] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Association for Computing Machinery*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[22] Mienye, I.D. and Sun, Y. (2022) A Survey of Ensemble Learning: Concepts, Algorithms, Applications and Prospects. *IEEE Access*, **10**, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

[23] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., *et al*. (2012) Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, **120**, 25-36. https://doi.org/10.1016/j.rse.2011.11.026

[24] Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., *et al*. (2017) Copernicus Sentinel-2A Calibration and Products Validation Status. *Remote Sensing*, **9**, Article No. 584. https://doi.org/10.3390/rs9060584

[25] Blaschke, T. (2010) Object Based Image Analysis for Remote Sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, **65**, 2-16. https://doi.org/10.1016/j.isprsjprs.2009.06.004

[26] Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., *et al*. (2014) Geographic Object-Based Image Analysis—Towards a New Paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, **87**, 180-191. https://doi.org/10.1016/j.isprsjprs.2013.09.014

[27] Neubert, P. and Protzel, P. (2014) Compact Watershed and Preemptive SLIC: On Improving Trade-offs of Superpixel Segmentation Algorithms. 2014 22*nd International Conference on Pattern Recognition*, Stockholm, 24-28 August 2014, 996-1001. https://doi.org/10.1109/ICPR.2014.181

[28] Vedaldi, A. and Soatto, S. (2008) Quick Shift and Kernel Methods for Mode Seeking. In: Forsyth, D., Torr, P. and Zisserman, A., Eds., *Computer Vision—ECCV* 2008. *ECCV* 2008. *Lecture Notes in Computer Science*, Vol. 5305, Springer, Berlin, 705-718. https://doi.org/10.1007/978-3-540-88693-8_52

[29] Cheng, Y. (1995) Mean Shift, Mode Seeking and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, 790-799. https://doi.org/10.1109/34.400568

[30] Rouse Jr., J.W., Haas, R.H., Schell, J.A. and Deering, D.W. (1974) Monitoring Ve-

getation Systems in the Great Plains with ERTS. In: Freden, S.C., Mercanti, E.P. and Becker, M.A., Eds., *Third Earth Resources Technology Satellite-1 Symposium*, NASA Special Publication, Washington DC, 309-313.

[31] Key, C.H. and Benson, N.C. (2006) Landscape Assessment: Ground Measure of Severity, the Composite Burn Index; and Remote Sensing of Severity, the Normalized Burn Ratio. USDA Forest Service, Rocky Mountain Research Station, Ogden.

[32] USGS (2017) Landsat Surface Reflectance-Derived Spectral Indices. Department of the Interior, United States Geological Survey, Washington DC.

[33] Trigg, S. and Flasse, S. (2001) An Evaluation of Different Bi-Spectral Spaces for Discriminating Burned Shrub-Savannah. *International Journal of Remote Sensing*, **22**, 2641-2647. https://doi.org/10.1080/01431160110053185

[34] Hardisky, M.A., Klemas, V. and Smart, R.M. (1983) The Influence of Soil Salinity Growth Form and Leaf Moisture on the Spectral Radiance of *Spartina alterniflora* Canopies. *Photogrammetric Engineering & Remote Sensing*, **49**, 77-83.

[35] Xie, L., Zhang, R., Zhan, J., Li, S., Shama, A., Zhan, R., *et al.* (2022) Wildfire Risk Assessment in Liangshan Prefecture, China Based on An Integration Machine Learning Algorithm. *Remote Sensing*, **14**, Article No. 4592. https://doi.org/10.3390/rs14184592

[36] Dong, H., Wu, H., Sun, P. and Ding, Y. (2022) Wildfire Prediction Model Based on Spatial and Temporal Characteristics: A Case Study of a Wildfire in Portugal's Montesinho Natural Park. *Sustainability*, **14**, Article No. 10107. https://doi.org/10.3390/su141610107

[37] Arjasakusuma, S., Kusuma, S.S., Vetrita, Y., Prasasti, I. and Arief, R. (2022) Monthly Burned-Area Mapping Using Multi-Sensor Integration of Sentinel-1 and Sentinel-2 and Machine Learning: Case Study of 2019's Fire Events in South Sumatra Province, Indonesia. *Remote Sensing Applications: Society and Environment*, **27**, Article ID: 100790. https://doi.org/10.1016/j.rsase.2022.100790

[38] Scikit-Optimize: Sequential Model-Based Optimization in Python (n.d.). https://scikit-optimize.github.io/

[39] Tharwat, A. (2021) Classification Assessment Methods. *Applied Computing and Informatics*, **17**, 168-192. https://doi.org/10.1016/j.aci.2018.08.003

[40] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010