Automated detection of burned areas in Costa Rica: a first approach

Detección automatizada de áreas quemadas en Costa Rica: un primer acercamiento

Brayan Rodríguez-Delgado[1], Daniela Vargas-Sanabria[2], Heileen Aguilar-Arias[3], José Andrés Umaña-Ortiz[4], Andrés Segura-Castillo[5]

Rodríguez-Delgado, B; Vargas-Sanabria, D; Aguilar-Arias, H; Umaña-Ortiz, J.A; Segura-Castillo, A. Automated detection of burned areas in Costa Rica: a first approach. *Tecnología en Marcha*. Vol. 39 Nº especial sobre Inteligencia Artificial. Febrero, 2026. Pág. 229-249.

https://doi.org/10.18845/tm.v39i5.8504

1 Universidad Estatal a Distancia, Costa Rica.
brayan.rodriguez@uned.cr
https://orcid.org/0009-0005-2966-2379
2 Laboratorio de Investigación e Innovación Tecnológica, Universidad Estatal a Distancia, Costa Rica.
dvargass@uned.ac.cr
http://orcid.org/0000-0003-2483-4926
3 Laboratorio PRIAS, Centro Nacional de Alta Tecnología, San José, Costa Rica.
haguilar@cenat.ac.cr
https://orcid.org/0000-0001-5838-3225
4 Laboratorio PRIAS, Centro Nacional de Alta Tecnología, San José, Costa Rica.
aumana@cenat.ac.cr
https://orcid.org/0000-0002-3677-920X
5 Laboratorio de Investigación e Innovación Tecnológica, Universidad Estatal a Distancia, Costa Rica.
asegurac@uned.ac.cr
https://orcid.org/0000-0001-5647-1176

Keywords

Burned areas; detection; classification; machine learning; practical applications of AI; computational performance.

Abstract

In Costa Rica despite diverse studies carried out by wildfires, collection data still is arduous fieldwork due to geographical conditions, there are zones where accessibility conditions prevent data collections. Satellite images are tools useful to study different zones to detect burned areas or their scars, but processing data by researchers requires too much time due to the number of files that need to be analyzed. We propose in this paper a framework based on machine learning and spectral index analysis to help burned area detection with efficient computational performance. Selecting as our study area in the Guanacaste Conservation Area, we obtained data from Sentinel-2 mission; we could detect the most probable zones affected by wildfire. Although this is a first step in the prevention of wildfire in protected zones, our results demonstrate the potential to develop a future robust detecting system.

Palabras clave

Áreas quemadas; detección; clasificación; aprendizaje automático; aplicaciones prácticas de IA, rendimiento computacional.

Resumen

En Costa Rica a pesar de diversos estudios sobre incendios forestales, la recolección de datos todavía es un trabajo arduo debido a las condiciones geográficas, hay zonas donde las condiciones de accesibilidad dificultan la recolección. Las imágenes satelitales son herramientas útiles en el estudio de diferentes zonas para detectar áreas quemadas y sus cicatrices. Sin embargo, el procesamiento por parte de los investigadores suele requerir tiempo debido a la cantidad de archivos que son necesarios para realizar el análisis. En este artículo proponemos un marco de trabajo basado en aprendizaje automático e índices espectrales para ayudar en la detección de áreas quemadas con un rendimiento computacional eficiente. Seleccionamos el Área de Conservación de Guanacaste como área de estudio para descargar imágenes de Sentinel-2, pudimos detectar zonas posiblemente afectadas por incendios forestales. A pesar de ser un primer acercamiento en la prevención de incendios forestales, nuestros resultados evidencian el potencial para desarrollar un futuro sistema de detección robusto.

Introduction

Wildfires are unplanned fires that can occur due to natural causes such as lightning or volcanic activity, as well as, unauthorized or accidental human interventions [1]. They can have negative effects in several ways, for instance, forest and biodiversity loss, production of carbon emissions and other risky substances for human health [2], [3], [4]. Nowadays there are several statistical and visualization tools to study wildfires [5]. Digitalization and collection of information through satellite monitoring [6] and other remote sensing tools are extended among researchers [2], [4], [6].

Studying Burned Areas (BA) has vital importance to develop applications for mapping high risk areas in order to contribute to adequate prevention policies [1]. A burned area is a territorial surface with a certain degree of accumulation of ash and charcoal, removal or alteration of vegetation due wildfires [1].

On the other hand, Machine learning (ML) is a group of diverse computational tools widely used, together with multi-spectral indices analysis, to process satellite images for locating burned areas in different contexts [7]. Both ML supervised and unsupervised algorithms [7], [8], [9] have been utilized in detecting active fire, for instance, through neural networks to classify pre-processed image with spectral indices [8], visualization of land cover [7], and mapping burned areas [9].

In Costa Rica, different studies have dealt with the wildfires, especially at Guanacaste Conservation Area (GCA), in northwestern region; among most recently published work, we can mention the development of a model of probability of fire occurrence based on generalized Least Squares Model [10]; the proposal of a geomorphometric model to determine topographic parameter to controlling wildfires in tropical dry forest [11] and a multi-temporal analysis based in Burned Area Index to discriminate affected areas and severity of fires using satellite images [12], [13].

However, Campos-Vargas and Vargas-Sanabria indicated data collection on burned areas at GCA has involved arduous fieldwork [10]. Due to some geographical characteristics of Costa Rica, it is not always possible to access some regions to collect data *in situ* [10]. In addition, the consulting information of wildfires reported by GCA are restricted to their protection area; nevertheless, a greater amount of data is necessary to obtain a better comprehension of the wildfires dynamics.

Therefore, we explored the use of different automated algorithms that allow us to detect burning areas. Our main objective is to implement an efficient automated system capable of analyzing satellite images, from which to obtain these areas affected by wildfires to extract useful information for future research.

Material and methods

Our methodology begins by defining the study zone through geographic coordinates capturing a rectangle representing a zone on the ground. Then we download files called rasters with satellite sensor data containing the coordinates of this rectangle, and we crop all collected rasters centered in the defined area. After that, we normalize these rasters to reduce the number of outliers [14]. With normalized files, we calculate special metrics called Spectral Indexes, and the decision threshold for each index, using natural breaks algorithm, to classify the raster pixels in burned or unburned [15]. Nevertheless, Spectral Analysis could be insufficient resource to achieve a better result separating burned pixels from unburned ones, then we use Machine Learning methods to obtain the burned area excluding other non-affected pixels in raster images [7]. In the next sections we explain in detail every step.

Rasters

As we mentioned before, our main inputs are raster image files; they are data structures with one or more associated matrices where information is stored [16]. Each point (x,y) in the raster is matched with a geographic point (longitude and latitude) representing a small square in ground called pixel. In addition the pixel coordinates refers to an entry inside the matrix of values stored inside the file [16], this means for each image of n column and m rows, it will be necessary processing $n \times m$ points. In satellite imagery the information captured by sensors is stored in

rasters identified with a number that corresponds with a range of spectrum wavelength called band [17], [18], [19]. Before downloading a group of rasters, we defined the study area, as we explain in the next section.

Study Area

The study was carried out in the Guanacaste Conservation Area (GCA) located in the province of Guanacaste (Figure 1). The GCA is dominated by deciduous vegetation, which loses its foliage during the dry season from December to mid-May, when most wildfires occur [10] with approximately 180 days per year reaching temperatures close to 32°C [20].

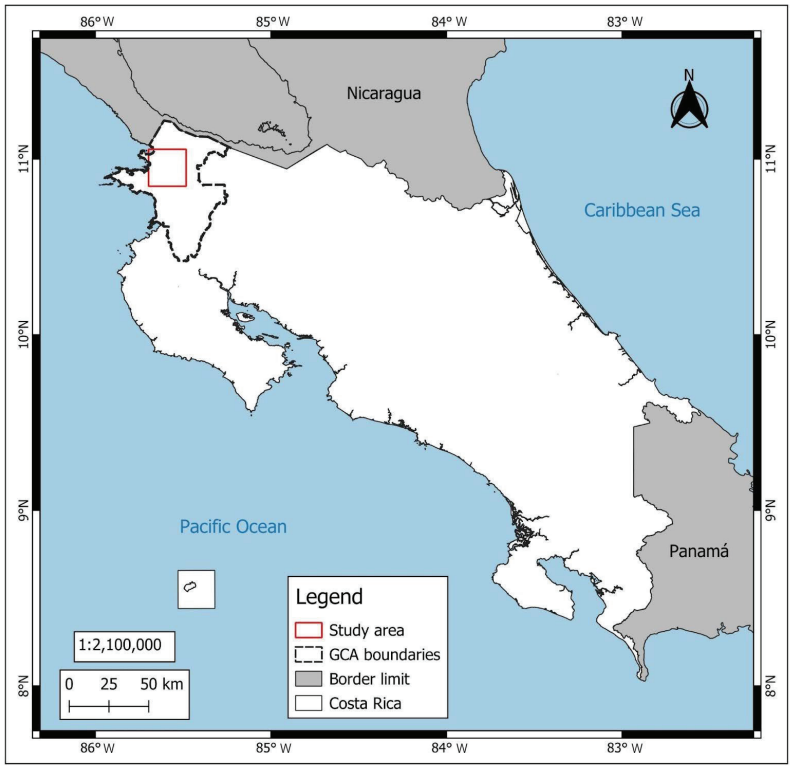

Figure 1. Guanacaste Conservation Area (GCA) study site.

The rest of the year, this region is characterized by significant rainfall during the rainy season [21] with precipitation between 900 to 2400 mm [11]. The area has different types of dry forest and a wide variety of coastal, riparian, and forest vegetation. However, some habitats have been transformed by illegal logging, wildfires, pasture planting and agricultural activities [11], [22]. For the test of our code implementation we focus on the zone defined by a rectangle between the geographic points (-85.697479, 10.845445) and (-85.697479, 11.055777). In 2023, there was the highest incidence of wildfires in GCA in the selected area, according to SINAC [23], [24]. We captured these rasters using study area geographic coordinates from the satellite mission Sentinel-2.

Earth Observation: Satellite Technologies

Sentinel-2 is a group of twin satellites (Sentinel-2A and Sentinel 2B) developed by the Copernicus Programme of the European Space Agency (ESA) to provide high resolution images that help monitor crops, forests, urban areas and water resources [25]. The free access policy, high accuracy with machine learning classifiers, and preprocessed data has resulted in an increasing use of Sentinel-2 images, especially in countries where access to satellite imagery is limited [25]. Sentinel-2 has a low geometric error [25]; its geolocation accuracy is better than 14.5 meters [26], while Landsat-8, another constellation developed by the National Aeronautics Space Administration (NASA), has a geolocation accuracy of 17 meters [27]. Besides, some of the products Sentinel-2 have geometric corrections [25] with a georeferencing error maximum value of 1.5 pixels [28]. These features are suitable for this research, therefore we selected a group of rasters between January - May 2023 and December 2023 - May 2024, corresponding to the Costa Rica dry season for the study site, when wildfire incidence is higher. We selected images with a cloud cover of 15% or less for testing our chosen algorithms [29]. These images have spatial resolutions of 10m, 20m, and 60m. For this study, the 20m was applied, i.e., one pixel in the image becomes 20 x 20 meters in the actual terrain.

Each group of Sentinel-2 images is compressed in files *JPG2000* (.jp2); these files can be converted to standard georeferenced files geotiff (.TIFF) denominated rasters [18]. The information in rasters can be read by the python library *rasterio*, storing data in a grid structure where each pixel of image is identified by a row-column coordinate system associated with their corresponding geographic coordinate. We preferred to use the name of the band or its acronym rather than the identifier number [15], equivalences are shown in Table 1.

Table 1. Sentinel-2 Band number identifier, name and acronym.

Band	Name	Band	Name
B01	Aerosol Coastal/Ultrablue (Ultrablue)	B07	Vegetation Red Edge3 (VRE3)
B02	Blue (BLUE)	B08	Near Infrared (NIR)
		B8A	Near Infrared Narrow (NIR_NARROW)
B03	Green (GREEN)	B09	Vegetation Red Edge4 (VRE4)
B04	Red (RED)	B10	Cirrus
B05	Vegetation Red Edge1 (VRE1)	B11	Shortwave Infrared1 (SWIR1)
B06	Vegetation Red Edge2 (VRE2)	B12	Shortwave Infrared2 (SWIR2)

Source: [14], [17], [19], [30]

We cropped the group images concentrating in the area with the highest number of fires reported by GCA [23], in order to speed up processing time. Before proceeding to obtain the spectral indices values for this area, we normalized band rasters files using Abrahim *et al.* method [14].

Band data normalization

The normalization process standardizes the data obtained from raster files to obtain results less susceptible to extreme values (outliers) [14]. We selected the percentile normalization following the method successfully applied by Abrahim *et al.* [46] to detect water bodies. We implemented a function following equation (1) with Python *numpy.clip* and *numpy.percentile* functions:

$$B_{normalized} = clip\left(\frac{B - P_2}{P_{98} - P_2}, 0, 1\right) \quad (1)$$

Where, all values in the raster band are converted into values from 0 to 1; setting the values less than percentile 2 to 0 and the ones over percentile 98 to 1. With normalized raster we obtained spectral inaaadices values.

Spectral Indices Analysis for burned area detection

For monitoring burned areas using satellite images it is common to start with spectral index analysis [15], [17], [18]. These indices are calculated using combinations of spectral values obtained from several bands [15], we explored a set of spectral indexes mentioned in reviewed literature [7], [15], [17], [18], [19], [31]. These indexes are summarized in Table 2. We use NIR_NARROW instead of NIR [17] because band 8 is omitted in the input Sentinel-2 rasters group (R20m). Thresholds are values that they allow to classify pixels in an affected or non-affected area [15].

Table 2. Spectral Index Summary.

Index	Equation, threshold and remarks
Normalized Burning Ratio (NBR) [17], [31]	$$NBR = \frac{SWIR2 - NIR}{SWIR2 + NIR} \quad (2)$$ Where values under 0.08 are considered burned [18].
NBR2 [32]	$$NBR2 = \frac{SWIR2 - SWIR1}{SWIR2 + SWIR1} \quad (3)$$ Dvorakova, *et al.* suggests a value of 0.05 as threshold [32].
NBRPLUS [17]	$$NBR+= \frac{SWIR2 - NIR_{narrow} - GREEN - BLUE}{SWIR2 + NIR_{narrow} + GREEN + BLUE} \quad (4)$$ Here, higher values represent burned areas [17]
NBRSWIR [17]	$$NBRSWIR = \frac{SWIR2 - SWIR1 - 0.02}{SWIR2 + SWIR1 + 0.1} \quad (5)$$
Normalized Difference Shortwave Infrared Index (NDSWIR) [17]	$$NDSWIR = \frac{SWIR1 - NIR_{narrow}}{SWIR1 + NIR_{narrow}} \quad (6)$$
Normalized Difference Vegetation Index (NDVI) [18], [33]	$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (7)$$
Green Normalized Difference Vegetation Index (GNDVI) [34], [35]	$$GNDVI = \frac{NIR - GREEN}{NIR + GREEN} \quad (8)$$ Values under 0.2 indicates spare or no vegetation [35]

Index	Equation, threshold and remarks
Enhanced Vegetation Index (EVI) [36], [37]	$$EVI = \frac{2.5 \cdot (NIR - RED)}{NIR + 6.0 \cdot RED - 7.5 \cdot BLUE + 1.0}$$ (9) This index could be affected in place near the equator [38]. We discarded this index from our analysis
Global Environmental Monitoring Index (GEMI) [36]	$$\eta = \frac{2 \cdot (NIR^2 - RED^2) + (1.5 + NIR) + (0.5 \cdot RED)}{NIR + RED + 0.5}$$ (10) $$GEMI = \frac{(1.0 - 0.25 * \eta) - (RED - 0.125)}{(1 - RED)}$$ (11) This is not recommended for dry or spare vegetation places [39]. Therefore we discarded this index for analysis.
Char Soil Index (CSI) [18]	$$CSI = \frac{NIR}{SWIR}$$ (12)
Analytical Burned Area Index (ABAI) [18]	$$ABAI = \frac{3 \cdot SWIR2 - 2 \cdot SWIR1 - 3 \cdot GREEN}{3 \cdot SWIR2 + 2 \cdot SWIR1 + 3 \cdot GREEN}$$ (13) For Guo *et al.* a burned area has values greater than -0.294 [18]
Burned Area Index (BAI) [36]	$$BAI = \frac{1}{(0.1 - RED)^2 + (0.06 - NIR)^2}$$ (14)
BAIM (Modis) [40]	$$BAIM = \frac{1}{(0.05 - NIR)^2 + (0.2 - SWIR)^2}$$ (15) We use BAIS2 instead of this index because we work with Sentinel-2 images.
BAIS2 (Sentinel-2) [17]	$$BAIS2 = \left(1 - \sqrt{\frac{VE2 \cdot VE3 \cdot NIR_{narrow}}{RED}}\right) \cdot \left(\frac{SWIR2 - NIR_{narrow}}{\sqrt{SWIR2 - NIR_{narrow}}} + 1\right)$$ (16) From -1 to 1 values represent a fire scar, and higher 1 values burned area [17]
Mid InfraRed Burn Index (MIRBI) [17], [36]	$$MIRBI = (10.0 \cdot SWIR2) - (9.8 \cdot SWIR1) + 2.0$$ (17)

Source: [17], [18], [33], [34], [35], [36], [37], [38], [39], [40]

On the other hand, sometimes, water bodies could interfere with the detection of burned areas, therefore data is usually filtered with Normalized Difference Water Index (NDWI) [17] (equation 18) or Water Pixels (WP) [41] (equation 19), where values greater than 0.0 represent water bodies, and they should be filtered through boolean operation with rasters [17].

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \quad (18)$$

$$WP = \frac{(NIR + SWIR1 + SWIR2) - (Ultrablue + BLUE + GREEN)}{(NIR + SWIR1 + SWIR2) + (Ultrablue + BLUE + GREEN)} \quad (19)$$

To improve the processing of data, we also utilized a procedure to mitigate cloud effect in raster data values [42]. Finally, the values of indices can be stored in new raster files to use later in the training process of machine learning methods [7], [17]. Therefore it is required defining threshold calculation methods to automate labeling of pixels.

Threshold Calculating

In the threshold calculating, it is common to use Jenks-Fisher's algorithm called "natural breaks" [15], this method is implemented in Python by the *jenskpy* library. This algorithm calculates for each possible way of clustering a point into a data set the square deviation from general mean, then calculates the goodness variance fit (GVF) that is the difference between the unity and the proportion between square deviation mean and the sum of all squared deviation for each mean of each subset [43] as is shown in equation (20).

$$GVF = 1 - \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{\sum_{k=1}^{m} \sum_{i=i}^{N_k} (x_{nk} - \mu_k)^2} \quad (20)$$

Where μ is the general average, m is the number of classes, N is the total of elements of the set, and μ_k is the mean in each subset. The main problem with this method is the computer processing time; equation (21) shows the number of possible combinations.

$$nCm = \frac{n!}{m!\,(n-m)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-m-1)}{m!} \quad (21)$$

Where n (number of points) and m (number of classes). The algorithm tests all these combinations [43], i.e., with a larger n, Jenks-Fisher algorithm needs to be executed approximately n^m times. For a faster calculation of "breaks" we explore the use of an alternative algorithm: K-means clustering [44]. K-means starts from an initial partition in k classes of the dataset, then an average called centroid is calculated in each group. After that each point in the dataset is rearranged according to its proximity with one of the centroids; with the new arranged class, centroids are newly obtained and repeated all rearranging until to reach a convergence (stable groups) or a maximum of iterations [44]. K-means is also implemented in Python through scikit-learn library, with a complexity value of O(n*m*d*t) [45], here n is the number of points, m number of class (m < n), d the dimension (in this case d = 2) and t the number of iterations (usually 300), this means this algorithm works with a complexity less than $O(n^2)$, i.e., the time of computing using K-means is smaller than Jenks-Fisher algorithm. We use the elbow methods to find the optimal number of classes in K-means [46]. To compare results of K-means and Fisher-Jenks algorithms we use confidence intervals of means [47]. With threshold values we proceed to validate spectral index analysis results.

Spectral index thresholds validation

We validate the obtained thresholds from spectral indexes rasters with choosing a sample of values expecting a low margin of error m_e calculated with equation (22) adjusted by population [48]:

$$m_e = \pm z_\alpha \cdot \sqrt{\frac{p(1-p)(N-n)}{n(N-1)}} \quad (22)$$

Where z is a normal distribution value with an alpha of 0.05 (z = 1.96), *p* proportion of success waited, *n* size of sample and *N* population size.

After sample the raster we construct a confusion matrix and calculate the precision and accuracy metrics [49], [50]. The metrics used to evaluate our confusion matrices include overall accuracy (OA), which measures the proportion of correct predictions; user accuracy (UA) or precision, which assesses the accuracy of positive predictions; and recall or producer accuracy (PA), representing the ratio of correctly classified items to the total items (Table 3).

Table 3. Metrics for evaluating confusion Matrix.

Metric	Equations
OA	$OA = \frac{TP + TN}{FN + TP + FP + TN}$ (23)
UA	$UA = \frac{TP}{TP + FP}$ (24)
PA	$PA = \frac{TP}{TP + FN}$ (25)
F1 Score	$F1Score = \frac{(\beta^2 + 1) \cdot PA \cdot UA}{\beta^2 \cdot PA + UA}$ (26)

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative
Source: [49], [50].

The F1 Score metrics (equation 26) weights UA and PA importance with β value. Usually, β is set to 1 according to [49], [50]. The spectral indexes on their own cannot assure an optimal automated burned area finding [7]. Nevertheless, they can be used as input for Automated ML Classifiers [7]. We study different algorithms of ML to filter burned areas in the next section to select those algorithms with better performances.

Selecting Machine Learning Classifiers

Several models of Artificial Intelligence (AI) including Machine Learning (ML) algorithms have been used in fire and burned area monitoring. Among ML models k-Nearest Neighbour (kNN), Random Forest (RF) are mentioned in [51], [52]; also Logistic Regression (LR) has been utilized as a classifier [53]. The kNN is based on calculating the proximity of a value and its centroid by means of its distance, and recalculating centroids to find stable groups similarly to k-means but with a pre-labeled data set [12], [31], [51]. On the other hand, RF uses decision tree structures with thresholds defined by rules of “greater than”, “less than” or “equal to” for classifying values [51].

We can also using the Naive Bayes (NB) classifier that uses Bayes Probability Theorem to convert a conditional relation between two variables into an inverse conditional relation [7], [54], i.e., we can pass from a probability of a value of spectral index given a burned pixel to the probability to find a burned pixel given spectral index value and classifying the pixel under this criterion. As we work classified values in two classes burned and unburned, we can take advantage of Support

Vector Machine (SVM) which uses a decision boundary (usually an hyperplane) to maximize the distance between subsets [55]. SVM, kNN are slow classifiers [56], RF training time could be slower depending on the amount of trees on it [57], faster alternatives are Extreme Gradient Boosting (XGB) and Histogram Gradient Boosting (HGB), both decision tree based classifiers [58], [59], [60]. We selected fast ML methods: NB, XGB and HGB for comparing classification against Spectral Index results, we also use LR algorithm as comparison point [59]. For finding the best parameters (*tuning*) of ML models, we use *RandomSearch* function [58].

All these algorithms need a pre-labeled set to be trained because they are supervised models [61], [62], we develop an automatic labeling process taking as input the decision threshold. We discuss the method in the next section.

Data Labeling

Raster file data structures are essentially matrices [18], where each entry represents a pixel with an assigned value. Therefore, we can apply linear mapping to assign a value of 0 or 1 to each pixel [9] representing unburned or burned respectively [8]. So, let $X \in IR^{mxn}$ be a matrix with values of any spectral index for each pixel in a raster file, we can define $v: IR \rightarrow \{0, 1\}$ as shown in equation (27), for each x_{ij} in X:

$$v(x_{ij}) = \begin{cases} 1 \ if\ value\ fulfill\ a\ certain\ condition \\ 0 \ in\ other\ cases \end{cases} \quad (27)$$

Then, we obtain a binary matrix $A \in IR^{mxn}$, where each a_{ij} entry is equal to $v(x_{ij})$. We define the condition: if the value x_{ij} is less (or greater) than a threshold value according to specific spectral indexes [17], [18]. This function was implemented using features of the *numpy* library. The values of the spectral index matrix and its associated binary of labels matrix will eventually be used to train the ML Classifiers.

Training and Evaluation of Machine Learning Classifications Methods

For training of the classification algorithms, samples were splitted using rule 70-30 (70% for training set and 30% for testing) [63]. Then, we constructed a confusion matrix to obtain evaluation metrics [50].

For evaluating ML models we also can include among the metrics: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R-squared score (R^2) [53], [64], shown in table 4.

Table 4. Error Analysis Metrics.

Metric	Value
MAE	$MAE = \frac{1}{n}\sum \ \lvert y - \hat{y}\rvert$ (28)
MSE	$MSE = \frac{1}{n}\sum \ (y - \hat{y})^2$ (29)
RMSE	$RMSE = \sqrt{MSE}$ (30)
R^2	$R^2 = 1 - \frac{\sum \ (y - \hat{y})^2}{\sum \ \left(y - mean(y)\right)^2}$ (31)

n: number of values, y: true values, ŷ: predicted values
Source: [53], [64]

The best fit in the model is indicated with lower values of MAE, MSE and RMSE, also with higher values in R^2 near or equal to 1, a R^2 negative indicates a poor fit [64]. We use implemented error metrics in library *scikit-learn*. In our case, MAE and MSE will be equal due our labeling data system, because of $(\pm1)^2 = |\pm1| = 1$ and $0^2 = |0| = 0$. Then RMSE and R^2 will be the best option to watch the performance of selected ML models.

A common problem in training model is the generalization degree on data: underfitting (poor accuracy in sets), overfitting (low error rate in training data in train set, and a poor performance in test set, i. e., high error rate) [65] or the perfect fitting [66] could happen while in processing of data. Another possible issue is the imbalanced data (the number of elements of one class overpass excessively to another) and it might generate high values in accuracy or precision metrics [67], when this situation happens, the data must be resampled. Jouloudari *et al.* [68] suggest using Synthetic Minority Over-sampling Technique (SMOTE), in case of imbalanced data, Jaccard index (IoU) and ROC-AUC (Receiver Operating Characteristics Area under the curve) score are preferred to other accuracy metrics [37]. To balance our data set we use an implemented Python library called *imblearn* to generate more values for minority class [68].

Finally, we developed a wrapper library in python to integrate different components to apply our analysis in selected rasters in order to filter burned areas in our study zone. We present our results in the next section.

Results

Regarding the threshold calculation, for the K-means method we can use 3 or 4 classes (Figure 2), we set the algorithm in 4 classes to obtain the greater number of approximations to limits of class [46]. We use the same number of classes in Jenks-Fisher algorithms to compare results with confidence intervals.

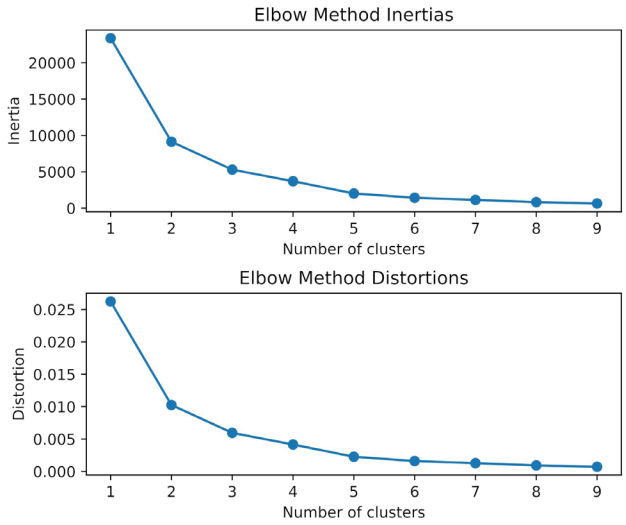

Figure 2. Elbow method to get optimum number of classes for K-means.

Respecting computation time processing, the Jenks-Fisher method took, on average, 1.41 hours by spectral index layer in raster (the complete process lasted 26.93 hours by raster), while K-means algorithm took 8 minutes by raster. Using the Kruskal-Wallis's test which could prove equality of variances between calculated breaks values (critical values < 0.11, p-value > 0.05), then we proceed to test means by confidence intervals (Table 5). With respect to Jenks-

Fisher and K-means algorithms to calculate threshold, in the analyzed index there was not found significant statistical evidence to affirm that there are differences between means, this means we can also use the faster K-means to estimate decision threshold for selected indexes [47].

Table 5. Confidence intervals from means of breaks values*.

Index	Jenks Fisher CI (95%)		K-Means CI (95%)	
	LI	LU	LI	LU
BAIS2	-2.67	2.47	-2.66	2.48
NBR	-2.64	2.50	-2.67	2.47
GNDVI	-2.15	2.99	-2.15	2.99
BAI	38.06	43.20	36.48	41.62
NBRSWIR	-2.41	2.73	-2.40	2.74
NDSWIR	-2.31	2.83	-2.27	2.87
NBR2	-2.63	2.51	-2.69	2.45
ABAI	-2.65	2.49	-2.69	2.45
MIRBI	-0.54	4.60	-0.45	4.69
NBRPLUS	-2.87	2.27	-3.00	2.14
NDVI	-2.03	3.11	-2.03	3.11
CSI	339.12	344.26	339.13	344.27

*Alpha = 0.05

In order to validate the performance of best spectral indices, we use reported burned zones of GCA in our study area, also we use VIIRS (Visible Infrared Imaging Radiometer Suite) at 375 meters of spatial resolution and MODIS (Moderate Resolution Imaging Spectroradiometer) at 1 kilometer spatial resolution points from the web site of FIRMS, and fires on vegetation attended and reported by the Fire Department of Costa Rica (Bomberos de Costa Rica) in situ as validation points.

With validation points we sample rasters with spectral index results to construct a confusion matrix and calculate the metrics of accuracy and precision (Table 6). We use *qgis* software to extract 2889 points (2184 burned, 705 unburned) between 2023 and 2024, with a confidence of 95% we estimate the margin of error in ± 1.82 (with a success proportion of 0.50). We discarded indexes with accuracy metrics below than ABAI values.

Table 6. Validation for best spectral indexes according to validation points.

index	Thres.	Confusion Matrix				Accuracy and Precision Metrics (%)			
		tp	tn	fp	fn	OA	UA	PA	F1 Score
MIRBI	> 3.487	1800	701	0	384	0.867	1.00	0.824	0.904
NBR2	< 0.216	2175	178	523	9	0.816	0.806	0.996	0.891
NBR	< -0.500	1403	701	0	781	0.729	1.00	0.642	0.782
ABAI	> 0.312	1129	701	0	1055	0.634	1.00	0.517	0.682

tp = True positive, tn = True negative, fp = False positive, fn = false negative, OA= overall accuracy, UA = User accuracy, PA = producer accuracy

Figures 3 and 4 show a correspondence between field data of GCA and burned area detected by Spectral Index analysis, but still it is required a better filtering of pixels by expert users.

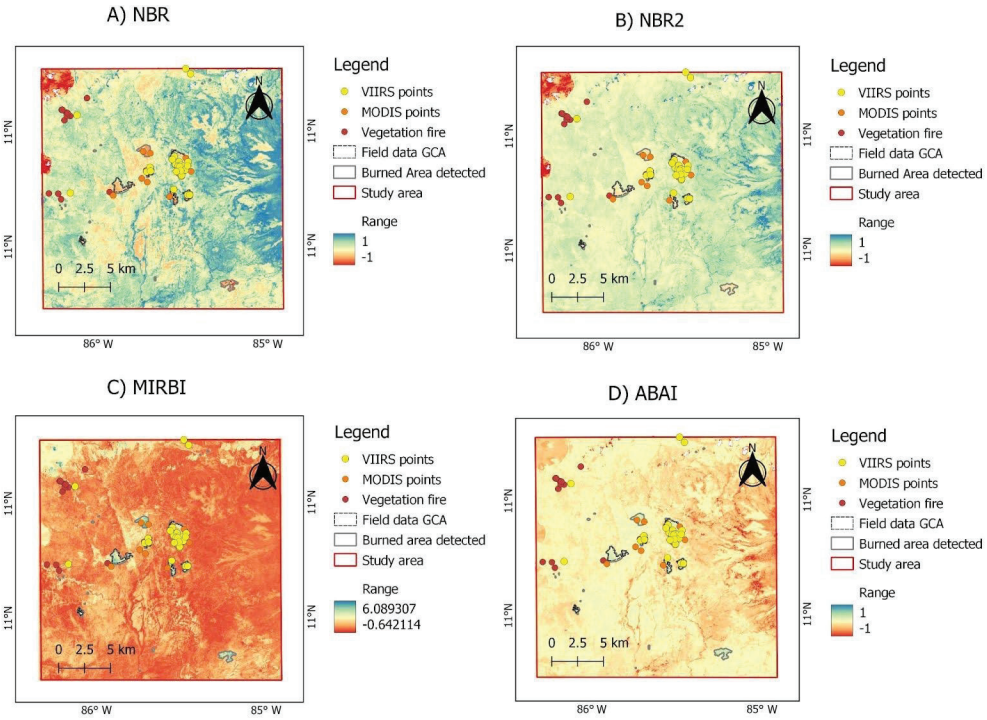

Figure 3. Validation of Spectral indices for the year 2023.

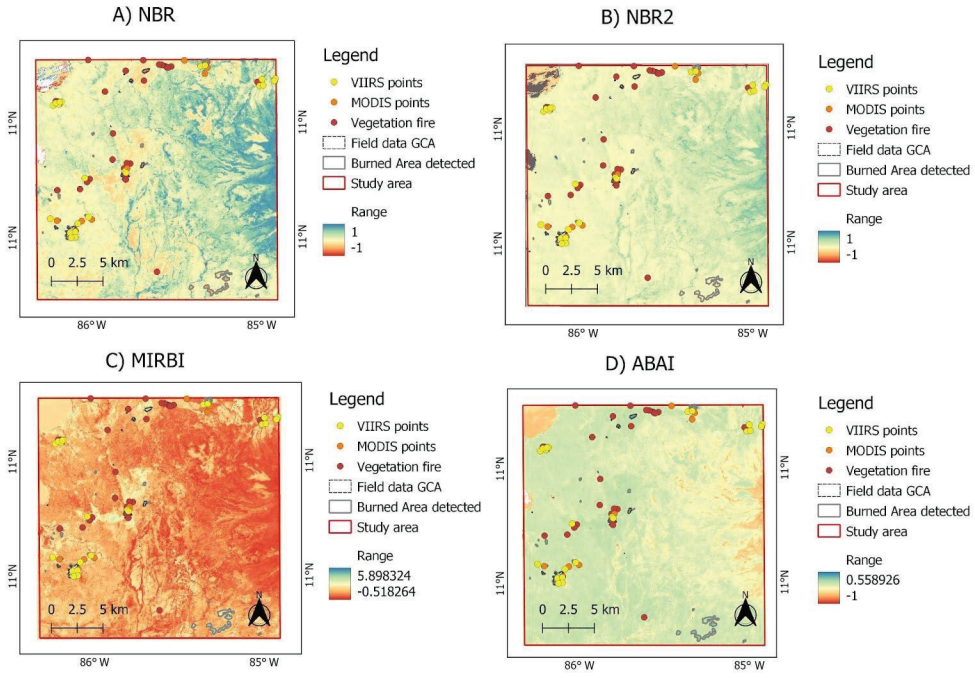

Figure 4. Validation of Spectral indices for the year 2024.

By selecting the more functional indexes, i.e., those which indicate where there was a burned area or scar of burned area, [17], [18], [31], [32], we proceed with a further analysis with ML Classifiers.

We trained models for LR, NB, and XGB, using our labeling method. The first step was finding the best hyperparameters. We used a grid with different values and a scikit-learn *RandomSearch* function. Then we trained models using the 70-30 rule for partitioning sets of rasters values [49]. The quicker method was NB (average < 15 seconds in training each Spectral Index raster set), followed by XGB (tuning: 563 s., training: 45 s.). The slower ones were HGB (tuning: 1137 s., training: 106s) and LR (tuning: 945 s., training: 85 s.). Accuracy and precision metrics were summarized in table 7. Best results, generally, were obtained with XGB and LR algorithms. Respecting the XGB model, similar results were obtained by Militino *et al.* [53].

We compared these models with a dummy model accuracy values with different parameter implemented in scikit-learn dummy test model (on average, stratified: 0.9907 ± 0.006, most_ frequent: 0.9952 ± 0.003, prior: 0.9952 ± 0.003, uniform: 0.50 ± 0.0001) [69], in general, the results of our selected ML models were better than the dummy model except MIRBI with NB model.

Table 7. LR and ML models according to Spectral Indexes.

SI	model	oa	pa	ua	f1	iou	k	roc_auc
NBR	LR	0.9998	0.9996	0.9996	0.9998	0.9996	0.9996	0.9998
	NB	0.9992	0.9984	0.9984	0.9992	0.9984	0.9984	0.9992
	XGB	0.9998	0.9996	0.9996	0.9998	0.9996	0.9996	0.9998
	HGB	0.9987	1.0000	1.0000	0.9987	0.9975	0.9975	0.9987
NBR2	LR	0.9997	0.9994	0.9994	0.9997	0.9994	0.9994	0.9997
	NB	0.9982	0.9964	0.9964	0.9982	0.9964	0.9964	0.9982
	XGB	0.9988	1.0000	1.0000	0.9988	0.9976	0.9976	0.9988
	HGB	0.9981	0.9963	0.9963	0.9981	0.9963	0.9963	0.9981
ABAI	LR	0.9998	0.9996	0.9996	0.9998	0.9996	0.9996	0.9998
	NB	0.9965	0.9931	0.9930	0.9965	0.9930	0.9931	0.9965
	XGB	0.9985	0.9969	0.9969	0.9985	0.9969	0.9969	0.9985
	HGB	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
MIRBI	LR	0.9997	0.9994	0.9994	0.9997	0.9994	0.9994	0.9997
	NB	0.9914	0.9830	0.9827	0.9914	0.9827	0.9830	0.9914
	XGB	0.9998	0.9995	0.9995	0.9998	0.9995	0.9995	0.9998
	HGB	0.9989	0.9978	0.9978	0.9989	0.9978	0.9978	0.9989

SI: Spectral Index, OA = Overall Accuracy, UA = User Accuracy, PA = Producer Accuracy, F1 = F1 Score, IOU = Jaccard Index, k = Cohen Kappa, ROC_AUC = Receiver Operating Characteristic Area Under the Curve.

Values in Error analysis indicates good fit of classifiers models [21]. The best models were LR, XGB and HGB (Table 8). Respecting XGB and HGB, differences in RMSE between are minimal; MAE, MSE and R2 are equal. NB had the lowest fitting performance. LR had the best performance but is a slower training method than XGB.

Table 8. Error Analysis of ML and LR models*.

Model	MAE and MSE	RMSE	R^2
LR	0.0003 ± 0.00007	0.016 ± 0.021	0.9990 ± 0.0003
NB	0.0037 ± 0.003	0.056 ± 0.028	0.9543 ± 0.013
XGB	0.0007 ± 0.0005	0.014 ± 0.013	0.9957 ± 0.002
HGB	0.0007 ± 0.0005	0.014 ± 0.016	0.9957 ± 0.004

*Average values

Despite the high values of performance metrics could imply a high risk of overfitting, we must consider the limited number of rasters with a low cloud covering percentage, which allows us to obtain comparable spectral index values during the wildfire season in the study area. Nevertheless we used the normalization of input files to reduce the incidence of outliers [14] and performed cross validation with available GCA information. Figure 5 shows how the ML models help to filter rasters with better precision, although still it is necessary for an expert interpretation; the most probable zone affected by a fire can be observed clearly and they closely match the

data reported by CGA. Great water bodies as ocean still are detected, but this could be filtered nevertheless, our method using ML together with Spectral Index is more effective than using only Spectral Index analysis [7] to distinguish burned areas, showing more clearly affected zones by fires. The best results were obtained with LR and boosting methods (XGB and HGB).

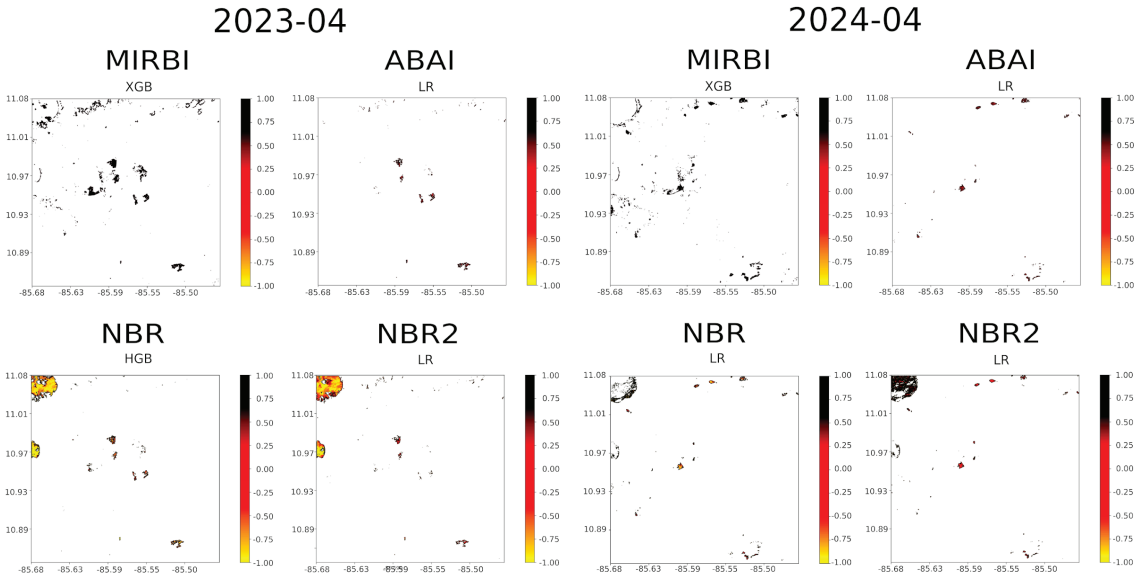

Figure 5. Classified pixels with ML models.

Conclusions

Our study demonstrated the possibility to integrate different unconnected python libraries to construct a system capable of detecting burned areas in our country, using rasters files as input, with successful results. This approach constitutes a relevant methodological contribution, as it demonstrates the feasibility of developing flexible and reproducible solutions based on open source software for environmental monitoring. The use of Spectral Index Analysis can be complemented with ML algorithms [7], which opens the possibility of improving the process automation and reducing reliance on manual interpretation in later stages.

Calculating a threshold for spectral analysis with natural breaks based in K-means is a useful fast alternative to defining spectral index limits for any region without similar previous studies [7], [15]. However, the performance of this approach may be influenced by spectral variability associated with seasonality, weather, and land cover, which suggests the needs for future comparative evaluations with other methods.

Our method was effective and quite efficient to detect burned zones in our study zone including scars of fires on lands, but still require expert perspective to interpret generated rasters. Nevertheless, our results indicate this method is precise enough without sacrificing computational performance and processing time. For future work, the incorporation of field data together with automated validation techniques is recommended, as these approaches may contribute to reducing subjectivity and enhancing the robustness and reliability of the results.

Although we focused on CGA, the approach can be used in different parts of Costa Rica for collecting information [12] through satellite imagery processing on high risk restricted or geographical difficult to access areas or other zones without previous historical information in order to map burned areas to understand damage extension, etc. [70]. Expanding the study to different regions would allow for the assessment of method's generalization capability and the adjustment of analysis parameters to diverse environmental and geographic contexts.

Finally, the generation of data on burned areas will allow the study of fire dynamics in sites affected by recurrent fires [11]. It will also support national strategies based on integrated fire management and develop potential conservation strategies [71]. Our research is one effort in the development of tools for monitoring and prevention of wildfires in Costa Rica and lays the groundwork for the future development of operational monitoring and early warning systems based on satellite image analysis.

Acknowledgments

This research was possible thanks to funding by CeNAT-CONARE Scholarship Program 2023-24, Costa Rica; with a special acknowledgment to Laboratory PRIAS-CeNAT, LIIT-UNED and ProRED-UNED for the support in data recollection, programming Python notebooks and scripts needed for this work.

References

[1] C. Maffei and M. Menenti, "Predicting forest fires burned area and rate of spread from pre-fire multispectral satellite measurements," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 263–278, Dec. 2019, doi: https://doi.org/10.1016/j.isprsjprs.2019.10.013.

[2] S. T. Seydi, M. Hasanlou, and J. Chanussot, "Burnt-Net: Wildfire burned area mapping with single post-fire Sentinel-2 data and deep learning morphological neural network," *Ecological Indicators*, vol. 140, p. 108999, Jul. 2022, doi: https://doi.org/10.1016/j.ecolind.2022.108999.

[3] X. Wang *et al.*, "Estimation of Forest Fire Burned Area by Distinguishing Non-Photosynthetic and Photosynthetic Vegetation Using Triangular Space Method," *Remote Sensing*, vol. 15, no. 12, p. 3115, Jun. 2023, doi: https://doi.org/10.3390/rs15123115.

[4] L. Jiao and Y. Bo, "Near real-time mapping of burned area by synergizing multiple satellites remote-sensing data," *GIScience & remote sensing/GIScience and remote sensing*, vol. 59, no. 1, pp. 1956–1977, Nov. 2022, doi: https://doi.org/10.1080/15481603.2022.2143690.

[5] S. Lakhina, *et al.* "Increasing National Resilience through an Open Disaster Data Initiative - Federation of American Scientists," *Federation of American Scientists*, Sep. 26, 2024. https://fas.org/publication/increasing-national-resilience-through-an-open-disaster-data-initiative/

[6] Shefali Juneja Lakhina and Anukool Lakhina, "Technology Entrepreneurship and Wildfire Risk Management," *Disaster risk reduction*, pp. 277–295, Jan. 2022, doi: https://doi.org/10.1007/978-981-19-2053-0_15.

[7] N. Chucos, & E. Vega. "Vista de Evaluación de algoritmos de machine learning en la clasificación de imágenes satelitales multiespectrales, caso: Amazonia Peruana," *Doi.org*, 2025. https://doi.org/10.37811/cl_rcm.v6i1.1843 (accessed Mar. 19, 2025)

[8] G. H. de Almeida Pereira, A. M. Fusioka, B. T. Nassu, and R. Minetto, "Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 171–186, Aug. 2021, doi: https://doi.org/10.1016/j.isprsjprs.2021.06.002.

[9] R. G. Negri, A. E. O. Luz, A. C. Frery, and W. Casaca, "Mapping Burned Areas with Multitemporal–Multispectral Data and Probabilistic Unsupervised Learning," *Remote Sensing*, vol. 14, no. 21, p. 5413, Jan. 2022, doi: https://doi.org/10.3390/rs142154

[10] C. Campos-Vargas and D. Vargas-Sanabria, "Assessing the probability of wildfire occurrences in a neotropical dry forest," *Écoscience*, vol. 28, no. 2, pp. 159–169, Apr. 2021, doi: https://doi.org/10.1080/11956860.2021.1916213.

[11] A. Quesada-Román and D. Vargas-Sanabria, "A geomorphometric model to determine topographic parameters controlling wildfires occurrence in tropical dry forests," *Journal of Arid Environments*, vol. 198, p. 104674, Mar. 2022, doi: https://doi.org/10.1016/j.jaridenv.2021.104674.

[12] D. Vargas-Sanabria and C. Campos-Vargas, "Comparación multitemporal de áreas quemadas en un bosque seco tropical utilizando el índice de área quemada (IAQ)," *Revista Forestal Mesoamericana Kurú*, vol. 17, no. 41, pp. 29–36, Jul. 2020, doi: https://doi.org/10.18845/rfmk.v17i41.5280.

[13] A. da, Antonio, A. M. Ruiz-Armenteros, R. F. Henriques, and Santos, "Analysis of Spectral Separability for Detecting Burned Areas Using Landsat-8 OLI/TIRS Images under Different Biomes in Brazil and Portugal," *Forests*, vol. 14, no. 4, pp. 663–663, Mar. 2023, doi: https://doi.org/10.3390/f14040663.

[14] B. I. Abrahim, Ammar Abd Jasim, M. R. Mahmood, H. R. Mahmood, H. A. Alalwan, and M. M. Mohammed, "Water Body Detection Using Sentinel-2 Imagery Through Particle Swarm Intelligence: A Novel Framework for Optimizing Spectral Multi-Band Index," *Eng—Advances in Engineering*, vol. 6, no. 3, pp. 59–59, Mar. 2025, doi: https://doi.org/10.3390/eng6030059.

[15] A. G. Flores Rodríguez, J. G. Flores-Garnica, D. R. González-Eguiarte, A. Gallegos-Rodríguez, P. Zarazúa-Villaseñor, and S. Mena-Munguía, "Análisis comparativo de índices espectrales para ubicar y dimensionar niveles de severidad de incendios forestales," *Investigaciones Geográficas*, no. 106, Nov. 2021, doi: https://doi.org/10.14350/rig.60396.

[16] L. Wasser *et al.*, "Datacarpentry/Organization-Geospatial: Data Carpentry Introduction To Geospatial Concepts August 2018 Release," *OPAL (Open@LaTrobe) (La Trobe University)*, Aug. 2018, doi: https://doi.org/10.5281/zenodo.1404414.

[17] E. Alcaras, D. Costantino, F. Guastaferro, C. Parente, and M. Pepe, "Normalized Burn Ratio Plus (NBR+): A New Index for Sentinel-2 Imagery," *Remote Sensing*, vol. 14, no. 7, p. 1727, Apr. 2022, doi: https://doi.org/10.3390/rs14071727.

[18] R. Guo, J. Yan, H. Zheng, and B. Wu, "Assessment of the Analytic Burned Area Index for Forest Fire Severity Detection Using Sentinel and Landsat Data," *Fire*, vol. 7, no. 1, p. 19, Jan. 2024, doi: https://doi.org/10.3390/fire7010019.

[19] S. T. Seydi, M. Akhoondzadeh, M. Amani, and S. Mahdavi, "Wildfire Damage Assessment over Australia Using Sentinel-2 Imagery and MODIS Land Cover Product within the Google Earth Engine Cloud Platform," *Remote Sensing*, vol. 13, no. 2, p. 220, Jan. 2021, doi: https://doi.org/10.3390/rs13020220.

[20] D. H. Janzen and W. Hallwachs, "Área de Conservación Guanacaste, northwestern Costa Rica: Converting a tropical national park to conservation *via* biodevelopment," *Biotropica*, vol. 52, no. 6, pp. 1017–1029, Jan. 2020, doi: https://doi.org/10.1111/btp.12755.

[21] M. Ríos-Solano, A. M. Durán-Quesada, C. Birkel, H. G. Hidalgo, W. Cabos, and D. V. Sein, "Observed interannual variability and projected scenarios of drought in the Chorotega region, Costa Rica," *Atmósfera*, vol. 38, Mar. 2024, doi: https://doi.org/10.20937/atm.53295.

[22] D.H., Janzen and W. Hallwachs, "Biodiversity conservation history and future in Costa Rica: the Case of Area de Conservación Guanacaste (ACG)". In *Costa Rican Ecosystems*, Kappelle, M. Ed. Chicago: The University Of Chicago Press, 2016, pp. 290-342.

[23] ACG, "Temporada de incendios forestales 2023 Área de Conservación Guanacaste - Área de Conservación Guanacaste," *Área de Conservación Guanacaste*, 2023. https://www.acguanacaste.ac.cr/noticias/noticias-programa-de-proteccion-e-incendios/5823-temporada-de-incendios-forestales-2023-area-de-conservacion-guanacaste (accessed Set. 06, 2025).

[24] M. De Ambiente, E. Sistema, N. De, and Á. De Conservación, "Informe anual Estadísticas SEMEC 2023 SINAC en Números." Available: https://www.sinac.go.cr/ES/transprncia/Informe%20SEMEC/Informe%20SEMEC%202023.pdf

[25] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 Data for Land Cover/Use Mapping: A Review," *Remote Sensing*, vol. 12, no. 14, p. 2291, Jul. 2020, doi: https://doi.org/10.3390/rs12142291.

[26] Rajagopalan Rengarajan, M. Choate, M. N. Hasan, and A. Denevan, "Co-registration accuracy between Landsat-8 and Sentinel-2 orthorectified products," *Remote Sensing of Environment*, vol. 301, pp. 113947–113947, Dec. 2023, doi: https://doi.org/10.1016/j.rse.2023.113947.

[27] L. Yan and D. P. Roy, "Improving Landsat Multispectral Scanner (MSS) geolocation by least-squares-adjustment based time-series co-registration," *Remote Sensing of Environment*, vol. 252, p. 112181, Jan. 2021, doi: https://doi.org/10.1016/j.rse.2020.112181.

[28] M. Vaqueiro, J. M. Fonseca, H. Oliveira, and A. Mora, "Multi-image Super-Resolution Algorithm Supported on Sentinel-2 Satellite Images Geolocation Error," vol. 53, pp. 50–57, Jul. 2021, doi: https://doi.org/10.1109/yef-ece52297.2021.9505092.

[29] H. Farhadi, M. Mokhtarzade, H. Ebadi, and B. A. Beirami, "Rapid and automatic burned area detection using sentinel-2 time-series images in google earth engine cloud platform: a case study over the Andika and Behbahan Regions, Iran," *Environmental Monitoring and Assessment*, vol. 194, no. 5, Apr. 2022, doi: https://doi.org/10.1007/s10661-022-10045-4.

[30] F. Ngadze, K. S. Mpakairi, B. Kavhu, H. Ndaimani, and M. S. Maremba, "Exploring the utility of Sentinel-2 MSI and Landsat 8 OLI in burned area mapping for a heterogenous savannah landscape," *PLOS ONE*, vol. 15, no. 5, p. e0232962, May 2020, doi: https://doi.org/10.1371/journal.pone.0232962.

[31] E. Kurbanov *et al.*, "Remote Sensing of Forest Burnt Area, Burn Severity, and Post-Fire Recovery: A Review," *Remote Sensing*, vol. 14, no. 19, p. 4714, Sep. 2022, doi: https://doi.org/10.3390/rs14194714.

[32] K. Dvorakova, U. Heiden, K. Pepers, G. Staats, Gera van Os, and Bas van Wesemael, "Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties," vol. 429, pp. 116128–116128, Nov. 2022, doi: https://doi.org/10.1016/j.geoderma.2022.116128.

[33] G. E. Suárez-Fernández, J. Martínez-Sánchez, and P. Arias, "Assessment of vegetation indices for mapping burned areas using a deep learning method and a comprehensive forest fire dataset from Landsat collection," *Advances in Space Research*, vol. 75, no. 2, pp. 1665–1685, Dec. 2024, doi: https://doi.org/10.1016/j.asr.2024.12.001

[34] Rabah Zennir and Boubaker Khallef, "Forest fire area detection using Sentinel-2 data: Case of the Beni Salah national forest – Algeria," *Journal of Forest Science*, vol. 69, no. 1, Jan. 2023, doi: https://doi.org/10.17221/50/2022-JFS.

[35] Shaparas Daliman, M. Jane, Pradnya Paramarta Raditya Rendra, Emi Sukiyah, Mohamad, and N. Sulaksana, "Dual Vegetation Index Analysis and Spatial Assessment in Kota Bharu, Kelantan using GIS and Remote Sensing," *BIO Web of Conferences*, vol. 131, pp. 05009–05009, Jan. 2024, doi: https://doi.org/10.1051/bioconf/202413105009.

[36] T. J. Hawbaker *et al.*, "The Landsat Burned Area algorithm and products for the conterminous United States," *Remote Sensing of Environment*, vol. 244, p. 111801, Jul. 2020, doi: https://doi.org/10.1016/j.rse.2020.111801.

[37] Hasan Bilgehan Makineci, "Investigation of burned areas with multiplatform remote sensing data on the Rhodes 2023 forest fires," *Ain Shams Engineering Journal*, pp. 102949–102949, Jul. 2024, doi: https://doi.org/10.1016/j.asej.2024.10294

[38] E. Rahimi, P. Dong, and C. Jung, "Global Pattern of Vegetation Homogeneity and Its Impact on Land Surface Temperature," *Land*, vol. 14, no. 2, pp. 421–421, Feb. 2025, doi: https://doi.org/10.3390/land14020421.

[39] Polina Lemenkova and Olivier Debeir, "Computing Vegetation Indices from the Satellite Images Using GRASS GIS Scripts for Monitoring Mangrove Forests in the Coastal Landscapes of Niger Delta, Nigeria," *Journal of Marine Science and Engineering*, vol. 11, no. 4, pp. 871–871, Apr. 2023, doi: https://doi.org/10.3390/jmse11040871

[40] M. P. Martín, I. Gómez, and E. Chuvieco, "Burnt Area Index (BAIM) for burned area discrimination at regional scale using MODIS data," *Forest Ecology and Management*, vol. 234, p. S221, Nov. 2006, doi: https://doi.org/10.1016/j.foreco.2006.08.248.

[41] Y. Zheng, L. Tang, and H. Wang, "An improved approach for monitoring urban built-up areas by combining NPP-VIIRS nighttime light, NDVI, NDWI, and NDBI," *Journal of Cleaner Production*, vol. 328, p. 129488, Dec. 2021, doi: https://doi.org/10.1016/j.jclepro.2021.129488.

[42] N. Zikiou, H. Rushmeier, M. I. Capel, Tarek Kandakji, N. Rios, and Mourad Lahdir, "Remote Sensing and Machine Learning for Accurate Fire Severity Mapping in Northern Algeria," *Remote Sensing*, vol. 16, no. 9, pp. 1517–1517, Apr. 2024, doi: https://doi.org/10.3390/rs16091517.

[43] M. Saleh, "Evaluation of Jenks Natural Breaks Clustering Algorithm for Changepoint Identification in Streaming Sensor Data," *IEEE Sensors Letters*, vol. 8, no. 10, pp. 1–4, Sep. 2024, doi: https://doi.org/10.1109/lsens.2024.3456292.

[44] M. Suyal and S. Sharma, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *Journal of Artificial Intelligence and Systems*, vol. 6, no. 1, pp. 85–95, Jan. 2024, doi: https://doi.org/10.33969/ais.2024060106.

[45] P. Singh Thakur, et. al. "Analysis of time complexity of K-means and fuzzy C-means clustering algorithm," *Engineering Mathematics Letters*, Jan. 2024, doi: https://doi.org/10.28919/eml/8402

[46] I. Herdiana, K. M. Alfin, Triyani, E. M. Nur, and Renny, "A More Precise Elbow Method for Optimum K-means Clustering," *arXiv (Cornell University)*, Feb. 2025, doi: https://doi.org/10.48550/arxiv.2502.008

[47] D. Chicco, A. Sichenze, and G. Jurman, "A simple guide to the use of Student's t-test, Mann-Whitney U test, Chi-squared test, and Kruskal-Wallis test in biostatistics," *BioData Mining*, vol. 18, no. 1, Aug. 2025, doi: https://doi.org/10.1186/s13040-025-00465-6

[48] S. K. Ahmed, "Research methodology simplified: How to choose the right sampling technique and determine the appropriate sample size for research," *Oral Oncology Reports*, vol. 12, no. 100662, pp. 1–7, 2024, doi: https://doi.org/10.1016/j.oor.2024.100662.

[49] J. Lu, W. Song, X. Zuo, D. Zhu, and Q. Wei, "A Study of Apple Orchards Extraction in the Zhaotong Region Based on Sentinel Images and Improved Spectral Angle Features," *Applied Sciences*, vol. 13, no. 20, p. 11194, Oct. 2023, doi: https://doi.org/10.3390/app132011194.

[50] I. El-Madafri, M. Peña, and N. Olmedo-Torre, "The Wildfire Dataset: Enhancing Deep Learning-Based Forest Fire Detection with a Diverse Evolving Open-Source Dataset Focused on Data Representativeness and a Novel Multi-Task Learning Approach," *Forests*, vol. 14, no. 9, p. 1697, Sep. 2023, doi: https://doi.org/10.3390/f14091697.

[51] A. da P. Pacheco, J. A. da S. Junior, A. M. Ruiz-Armenteros, and R. F. F. Henriques, "Assessment of k-Nearest Neighbor and Random Forest Classifiers for Mapping Forest Fire Areas in Central Portugal Using Landsat-8, Sentinel-2, and Terra Imagery," *Remote Sensing*, vol. 13, no. 7, p. 1345, Apr. 2021, doi: https://doi.org/10.3390/rs13071345.

[52] H. Xu *et al.*, "Immediate assessment of forest fire using a novel vegetation index and machine learning based on multi-platform, high temporal resolution remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 134, pp. 104210–104210, Oct. 2024, doi: https://doi.org/10.1016/j.jag.2024.104210.

[53] F. Chan, "Optimal Forecast Combination with Mean Absolute Error Loss," *CAMA Working Papers*, Nov. 2023. https://ideas.repec.org/p/een/camaaa/2023-59.html

[54] A. Ahmad *et al.*, "Naïve Bayes Classification of High-Resolution Aerial Imagery Optimization Modelling Analytic and Simulation (OptiMAS) Faculty of Information and Communication Technology," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021, Accessed: Nov. 06, 2025. [Online]. Available: https://thesai.org/Downloads/Volume12No11/Paper_20-Na%C3%AFve_Bayes_Classification_of_High_Resolution_Aerial_Imagery.pdf.

[55] F. Chollet, *Deep Learning with Python*. Shelter Island (New York, USA): Manning, Cop, 2018. Available: https://www.manning.com/books/deep-learning-with-python

[56] J. Cervantes, F. G. Lamont, L. R. Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, no. 1, pp. 189–215, Sep. 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[57] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102051, Jun. 2020, doi: https://doi.org/10.1016/j.jag.2020.102051.

[58] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, Oct. 2020, doi: https://doi.org/10.1016/j.jksuci.2020.10.013.

[59] A. F. Militino, H. Goyena, U. Pérez-Goya, and M. D. Ugarte, "Logistic regression versus XGBoost for detecting burned areas using satellite images," *Environmental and ecological statistics*, Jan. 2024, doi: https://doi.org/10.1007/s10651-023-00590-7.

[60] N. L. Fitriyani *et al.*, "A Novel Approach Utilizing Bagging, Histogram Gradient Boosting, and Advanced Feature Selection for Predicting the Onset of Cardiovascular Diseases," *Mathematics*, vol. 13, no. 13, pp. 2194–2194, Jul. 2025, doi: https://doi.org/10.3390/math13132194.

[61] A. Zhdanovskaya, D. Baidakova, and Dmitry Ustalov, "Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, pp. 15886–15893, Jun. 2023, doi: https://doi.org/10.1609/aaai.v37i13.26886.

[62] T. Bleckmann, and G. Friege, "Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method," *Knowledge Management & E-Learning: An International Journal*, pp. 433–447, Sep. 2023, doi: https://doi.org/10.34105/j.kmel.2023.15.025.

[63] V. Sheth, U. Tripathi, and A. Sharma, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," *Procedia Computer Science*, vol. 215, pp. 422–431, 2022, doi: https://doi.org/10.1016/j.procs.2022.12.044.

[64] K. Adil, Fayçal Messaoudi, A. Ahmed, and M. Youness, "Machine Learning and Deep Learning based Students' Grades Prediction," *Research Square (Research Square)*, Aug. 2023, doi: https://doi.org/10.21203/rs.3.rs-3192793/v1.

[65] C. Aliferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," *Computers in health care (New York)*, pp. 477–524, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-39355-6_10.

[66] Y. Sharon and D. Yehuda. "How Does Perfect Fitting Affect Representation Learning? On the Training Dynamics of Representations in Deep Neural Networks," *Arxiv.org*, 2021. https://arxiv.org/html/2405.17377v1#S5 (accessed Feb. 06, 2026).

[67] E. Jafarigol, T. Trafalis, and N. Mohammadi, "A Review of Machine Learning Techniques in Imbalanced Data and Future Trends," *arXiv.org*, 2023. https://arxiv.org/abs/2310.07917 (accessed Nov. 06, 2025).

[68] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences*, vol. 13, no. 6, p. 4006, Jan. 2023, doi: https://doi.org/10.3390/app13064006.

[69] M. A. Lones, "Avoiding common machine learning pitfalls," *Patterns*, p. 101046, Aug. 2024, doi: https://doi.org/10.1016/j.patter.2024.101046.

[70] P. Ettehadi Osgouei, E. Sertel, and M. E. Kabadayı, "Integrated usage of historical geospatial data and modern satellite images reveal long-term land use/cover changes in Bursa/Turkey, 1858–2020," *Scientific Reports*, vol. 12, no. 1, May 2022, doi: https://doi.org/10.1038/s41598-022-11396-1

[71] J. P. Ocón *et al.*, "Global tropical dry forest extent and cover: A comparative study of bioclimatic definitions using two climatic data sets," *PLOS ONE*, vol. 16, no. 5, p. e0252063, May 2021, doi: https://doi.org/10.1371/journal.pone.0252063.

Declaración sobre uso de Inteligencia Artificial (IA)

Los autores aquí firmantes declaramos que no se utilizó ninguna herramienta de IA para la conceptualización, traducción o redacción de este artículo.