DOI: 10.24850/j-tyca-14-04-03

Articles

Machine learning algorithms for classifying flood areas from synthetic aperture radar images

Algoritmos de aprendizaje automático para clasificar zonas de inundación a partir de imágenes de radar de apertura sintética

Juan Pablo Ambrosio-Ambrosio[1], ORCID: https://orcid.org/0000-0002-7032-6976

Juan Manuel González-Camacho[2], ORCID: https://orcid.org/0000-0001-5479-7316

[1]Colegio de Postgraduados Campus Montecillo, Montecillo, Mexico, ambrosio.juan@colpos.mx

[2]Colegio de Postgraduados Campus Montecillo, Montecillo, Mexico, jmgc@colpos.mx

Corresponding author: Juan Manuel González-Camacho, jmgc@colpos.mx

Abstract

The use of synthetic aperture radar (SAR) images represents a valuable source of information to characterize geographic regions susceptible to flooding, such as southeastern Mexico, they are not sensitive to cloudy and / or dark conditions. This research presents a methodology to identify bodies of water in a region of southeastern Mexico. Three machine learning algorithms were implemented: Random forests (RF), Gradient Boosting (GB) and Support Vector Machines (SVM) to classify three target classes: Class A (water, flooded areas, and bodies of water); class I (urban infrastructure and / or bare soil), and class V (vegetation) from SAR images. The SAR image used covers a projected geographical area UTM Zona 15 Norte WGS84 located in the states of Tabasco and Chiapas; this was pre-processed to reduce errors in the image. The RF, GB and SVM models were implemented in Python language. These were trained and tested in prediction from a database of 12 000 samples with amplitude values of the SAR image. The RF model obtained an overall classification accuracy (PG) of 97.9 (+/- 0.003) %; GB obtained PG = 97.9 (+/- 0.003) %, and SVM PG = 97.4 (+/- 0.005). The three models obtained an $F1\_s$ value higher than 0.99 to predict class A; RF obtained AUC = 1 for the three target classes. This study shows the potential use of SAR satellite images and the high performance of RF, GB and SVM machine learning models to classify and identify water bodies as well as highlighting its importance in studies of possible impacts of floods.

Keywords: Supervised learning, prediction models, satellite images, decision trees, watershed, remote sensing, ROC curves.

Resumen

El uso de imágenes de radar de apertura sintética (SAR) representa una fuente valiosa de información para caracterizar regiones geográficas susceptibles de inundaciones, como en el sureste de México, ya que éstas no son sensibles a condiciones de nubosidad y/u oscuridad. En esta investigación se presenta una metodología para identificar cuerpos de agua en una región del sureste de México. Se aplicaron tres algoritmos de aprendizaje automático: bosque aleatorio (RF), potenciación del gradiente (GB) y máquina de soporte vectorial (SVM) para clasificar las tres clases objetivo A: agua, áreas inundadas y cuerpos de agua; I: infraestructura urbana y/o suelo desnudo, y V: vegetación a partir de imágenes SAR. La imagen SAR utilizada cubre una zona geográfica proyectada UTM Zona 15 Norte WGS84, localizada en los estados de Tabasco y Chiapas, la cual fue preprocesada para disminuir errores en la imagen. Los modelos RF, GB y SVM se implementaron en lenguaje Python, que fueron entrenados y probados en predicción a partir de una base de datos de 12 000 muestras, con valores de amplitud de la imagen SAR. El modelo RF obtuvo una precisión global de clasificación (PG) de 0.979(+/-0.003); GB obtuvo PG = 0.979(+/-0.003), y SVM PG = 0.974(+/-0.005). Los tres modelos obtuvieron un valor de F1_*score* superior a 0.99 para predecir la clase A; el clasificador RF obtuvo valores de AUC = 1 para las

tres clases objetivo evaluadas. Este estudio permite mostrar el uso potencial de las imágenes satelitales SAR y el alto desempeño de los modelos de aprendizaje automático RF, GB y SVM para clasificar e identificar los cuerpos de agua, así como resaltar su importancia en estudios de los posibles impactos de las inundaciones.

Palabras clave: aprendizaje supervisado, modelos de predicción, imágenes satelitales, árboles de decisión, cuenca hidrológica, sensores remotos, curvas ROC.

Received: 12/22/2020

Accepted: 12/27/2021

Introduction

Today, synthetic aperture radar images obtained by remote sensors are available. Unlike optical sensors, radar images do not depend on reflected solar radiation or on heat radiation emitted by the earth. Rather, they emit their own electromagnetic radiation for their probes. Therefore, a

radar image is not affected by meteorological conditions or darkness (Sami & Abdulmunem, 2020; Fernández-Ordoñez & Soria-Ruiz, 2015). Satellite images cover large areas and even areas of difficult access. From their analysis, it is possible to extract information of interest indirectly. Processing radar images is a challenge in terms of dealing with deformations and noise caused by inclination of the sensor. However, overall accuracy of the classifier is conditioned by the level of processing of the data to reduce mottling in simple and multitemporal images, and by the characteristics of the images used as predictors, such as texture, color, object to be classified, agricultural areas, bodies of water, urban areas, and the number of target classes (Gomarasca *et al.*, 2019).

There are different approaches to classifying an image. Election of a specific approach depends on the nature of the object to be detected and the quantity and quality of the data. Machine learning algorithms have good predictive capacity, according to Avendaño-Pérez, Parra-Plazas and Fredy-Bayona (2014), who reported the use of support vector machine (SVM) with Gaussian kernel and a Bayesian model to classify water, soil, and population. They point out that performance of the SVM model was superior to the Bayesian model. Pulella, Aragão-Santos, Sica, Posovszky and Rizzoli (2020) mapped forested areas in the state of Rondonia, Brazil, with a random forest (RF) model trained with Sentinel-1 multitemporal radar images to classify artificial, forested, and non-forested areas; they obtained an overall classification accuracy of more than 80 %.

Radar images have been used in diverse fields of study. They have been used, for example, to identify oil spills, monitor sea ice, detect sea-going vehicles, classify ground cover, monitor soil moisture, and detect flooded areas. Shen, Wang, Mao, Anagnostou and Hong (2019) presented an ample review on the advantages and disadvantages of the use of these images to map flooded extensions. These authors reported acceptable accuracy in machine flood mapping with no obstructions. However, detection of floods under vegetation in urban areas was less satisfactory. Lin, Yun, Bhardwaj and Hill (2019) described a study for flood detection in urban areas from multitemporal Sentinel-1 satellite images after hurricane Matthew in 2016 in North Carolina, USA, and obtained low precision. In a similar manner, Zhang *et al*. (2018) reported the use of Sentinel-1ª and Sentinel-1B multitemporal data for zoning floods induced by hurricane Irma in Florida, USA, in 2017.

In southeastern Mexico periods of intense rains occur frequently in the months of August, September, October, and November, causing floods and economic losses in different sectors of the population and occasionally, losses of human life (Sánchez, Salcedo, Florido, & Mendoza, 2015).). For this reason, it is of interest to have reliable indirect methods of identifying areas affected by floods since in situ identification can be costly and slow. In these situations, satellite images and supervised machine learning models can be of great use in the classification of areas affected by floods.

The state of Tabasco must deal with floods because of geographic factors such as the presence of Mexico's two largest rivers (the Usumacinta and the Grijalva), mean annual precipitation of 2426 mm, lack of territorial planning, deforestation of the high parts of the watershed, climate change, and anthropic factors (Arreguín-Cortés & Rubio-Gutiérrez, 2014; Perevochtchikova & Lezama-de-la-Torre, 2010).

The objective of this study was to apply three machine learning algorithms (gradient boosting, random forest, and support vector machine) for zoning a geographic region in the states of Tabasco and Chiapas into three target classes: class A (flooded areas and bodies of water, class I (urban infrastructure and/or bare soil, and class V (vegetation), from synthetic aperture radar images. An additional objective was to illustrate the potential use of radar images to detect bodies of water. The radar image we analyzed corresponds to a scenario caused by the flood that occurred in the study area on October 8, 2017, in the state of Tabasco, Mexico.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Materials and methods

Study area

The study area is delimited by the red polygon seen in Figure 1, with reference to the World Geodetic System 1984 (WGS84) of geographic coordinates, with a projection to the UTM (Universal Transverse Mercator) system of coordinates. The synthetic aperture radar (SAR) image covers a large part of the state of Tabasco and a small part of the state of Chiapas. The study area includes 408 687 ha of this image.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

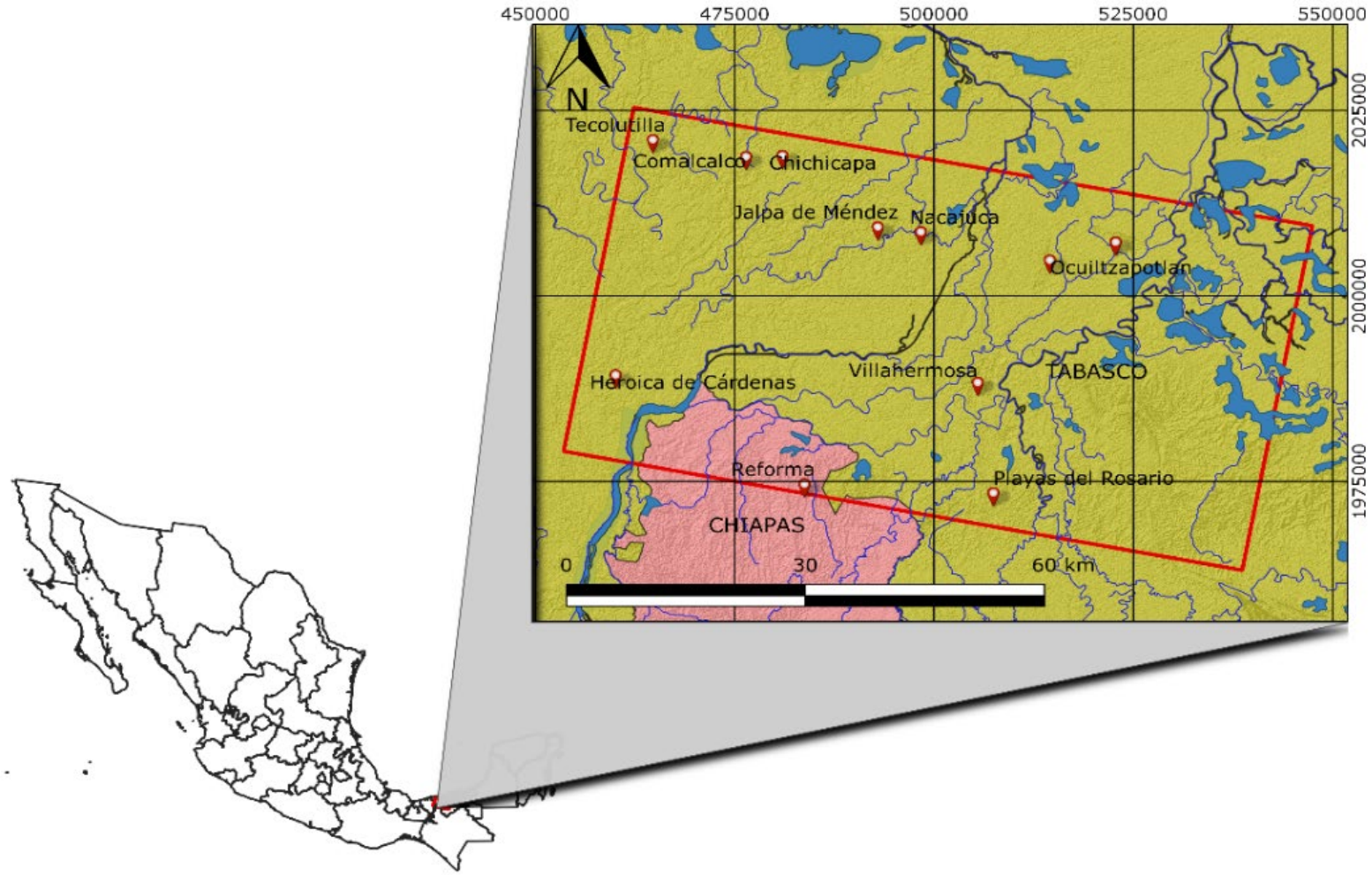

Figure 1. Study area, Tabasco and Chiapas, SRC UTM Zone 15 North WGS84.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Input data

To conduct this study, we obtained SAR images from the Sentinel-1A satellite. The images were downloaded through the WEB site of the Institute of Geophysics of the University of Alaska, Fairbanks (ASF, 2020). Table 1 presents the specific characteristics of the SAR image.

Table 1. Characteristics of the SAR image. Source: Copernicus Sentinel Data (2017).

Concept	Description
Product	S1A_IW_GRDH_1SDV_20171008T
Date of acquisition	8-OCT-2017 12:00:16.264908
Product level	1, Standard georeferenced product
Acquisition method	IW (Interferometric Wide)
Azimuth band width	327 Hz
Product type	GRD (Grand Range Detection)
Polarization	Dual VV + VH
Frequency	Band C
Step	Descendent

Digital processing of the images was performed with the free access software SNAP (sentinel application platform). The Sentinel 1 toolbox facilitates SAR image processing (ESA & SEOM, 2019). The SAR images were transformed to a matrix format with Matlab software (MathWorks

Inc., 2016), and geospatial processing of the SAR images was performed with QGIS software (QGIS.org, 2020). The machine learning models for classification were implemented in Python language with the Scikit-learn library (Pedregosa *et al*., 2011). Processing of images, data and models was carried out with a computational system in the 64-bit environment Windows 10, Intel Core i5 processer @2.50 GHz, and 8 GB RAM.

SAR image preprocessing

Prior to their analysis, the SAR images were preprocessed with SNAP. Figure 2 shows the preprocessing stages of the image.

Tecnología y Ciencias del Agua
OPEN ACCESS
Check for updates

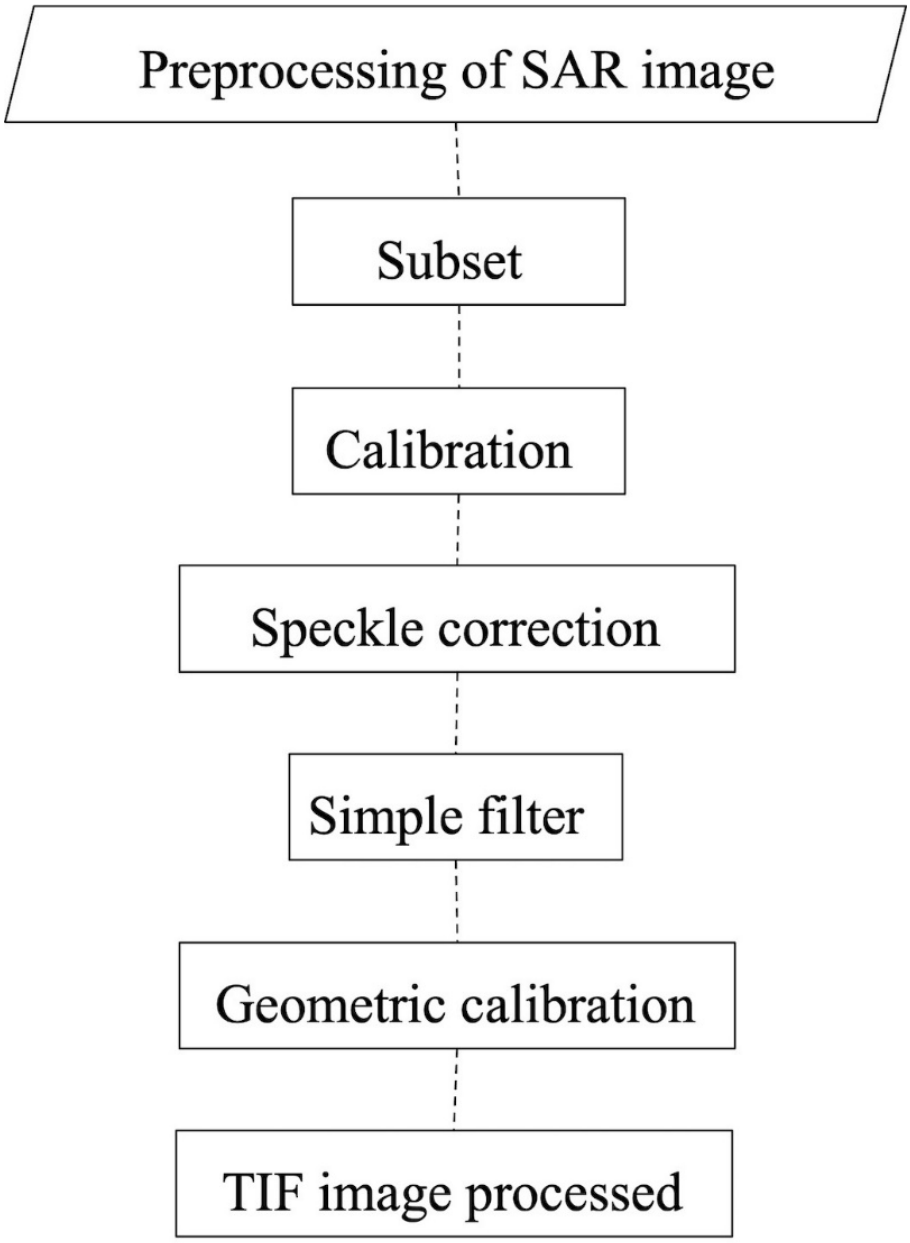

Figure 2. Preprocessing stages of the synthetic aperture radar (SAR) image.

The correction process consisted of defining a subsection of the original image that depends on the area of interest and the available computational processing capacity. To this subsection, radiometric correction was applied to eliminate the most visible mottles of the image; a Lee filter with parameters $look = 1$, and window size 3×3 is used. A geometric correction was then applied to georeference the image and facilitate its geographic visualization. Finally, the images were exported to a TIF graphic format (Abdurahman-Bayanudin & Heru-Jatmiko, 2016; Podest, 2018; UN-SPIDER, 2020).

Sample selection

The database was created based on samples of the SAR image associated with three target classes: class A (flooded areas and bodies of water), class I (urban infrastructure and/or bare soil, and class V (vegetation). The samples of the image highlight the characteristics of interest, such as rivers, lagoons, swamps, cities, highways, vegetation, and bare ground.

Class A comprises bodies of water: perennial rivers, intermittent rivers, lakes, lagoons, and flooded agricultural areas. The pixels of this

class were selected based on convenience sampling. The boundaries of the bodies of water were detected by two segmentation methods. The first, the regional growth method, consists of applying a dissimilarity parameter ($d = 0.035$) from a seed pixel to obtain a binary image with values equal to 1 for the pixels (x, y) similar to the seed and 0 otherwise. The second, the threshold method, consists of analyzing the histogram of the image to define the values of the upper and lower thresholds that segment the objects in the best possible way. This method generates a binary image with pixels (x, y) equal to 1 if they are found within the interval of the previously defined thresholds, and 0 otherwise.

Class A was created with the width values of the bands sigma0_VH and sigma0_VV for each pixel (x, y). These pixels were then exported to a text file with the coordinates (x, y) of each pixel and its sigma0_VH and sigma0_VV values.

Class I represents the physical infrastructure, which includes highways, buildings, houses, stores, factories, and soil with little or no vegetation that are visible in the Google satellite hybrid auxiliary image. To define this class, convenience sampling was used; to each sample of the image, binary segmentation was performed using a manually defined threshold. The sample of large cities that is illustrated in format *uint*8 was segmented with the threshold interval of $0.35 - 255$. The bands sigma0_VH and sigma0_VV were exported for the pixels (x, y) with values equal to 1, resulting in segmentation.

Class V represents vegetation: forests, grasslands, agricultural area, and jungle. This class is identified by means of a convenience sampling that consists of defining small windows that, in their totality, represent a type of vegetation. Then, by means of segmentation by regional growth ($d = 0.5$), the matrix of image data is accessed to extract the values of the bands sigma0_VH and sigma0_VV.

Machine learning algorithms

Three machine learning algorithms were selected for multiclass classification based on the characteristics of the SAR image bands.

Random forest model

The random forest model (RF) proposed by Breiman (2001) consists of a combination of predictors of decision trees in which each tree depends on

the values of a random independent vector with equal distribution for all the trees of the forest. Later, a large number of decision trees is generated, and the most frequently predicted class is selected. RF is defined as a classifier that comprises a set of structured tree classifiers $\{h(\mathbf{x}, \Theta k), k = 1, \ldots\}$, where Θk are independent random vectors, identically distributed, and each tree emits a unitary vote for the most frequent class, given an $\mathbf{x}$ entry. Geurts, Ernst and Wehenkel (2006), in the construction of a decision tree, proposed a replica of a learning sample bootstrap and the CART (classification and regression trees) algorithm together with the modification used in the sub-spatial method. In each test node, the optimal division is obtained by searching for a k-size subset of entry variable candidates (selected without replacement).

In the division of a node in a decision tree (in classification problems), the most used loss function is the index of Gini ($i_G(t)$) defined as (Gini, 1912):

$$L = i_G(t) = \sum_{k=1}^{J} p(C_k|t)\big(1 - p(C_k|t)\big) \qquad (1)$$

where $p(C_k|t)$ is the proportion of samples that belong to class k, and J represents the total number of classes for a particular node t.

Gradient boosting model

The gradient boosting model (GB) is defined by a set of decision trees that are trained sequentially using an iterative computational procedure where each subsequent model corrects the errors of its predecessor, resulting in a more robust model (Friedman, 2001). The loss function to be optimized is expressed as follows:

$$L(\{y_k, F_k(\mathbf{x})\}1^K) = -\sum_{k=1}^{K} y_k \log P_k(\mathbf{x}) \quad (2)$$

where $y_k \in \{0,1\};\ k = 1, \ldots, K$ classes, and $P_k(\mathbf{x}) = Pr(y_k = 1|\mathbf{x})$. Logistic regression (sigmoid function) is often used to find the probability that input $\mathbf{x}$ generates $y = 1$.

Support vector machine (SVM)

A support vector machine (SVM) is considered an extension of the perceptron model (Raschka & Mirjalili, 2017). SVM maximizes the margin,

that is, the distance between the separation hyperplane (decision limit) and the training samples next to this hyperplane called support vectors (Figure 3).

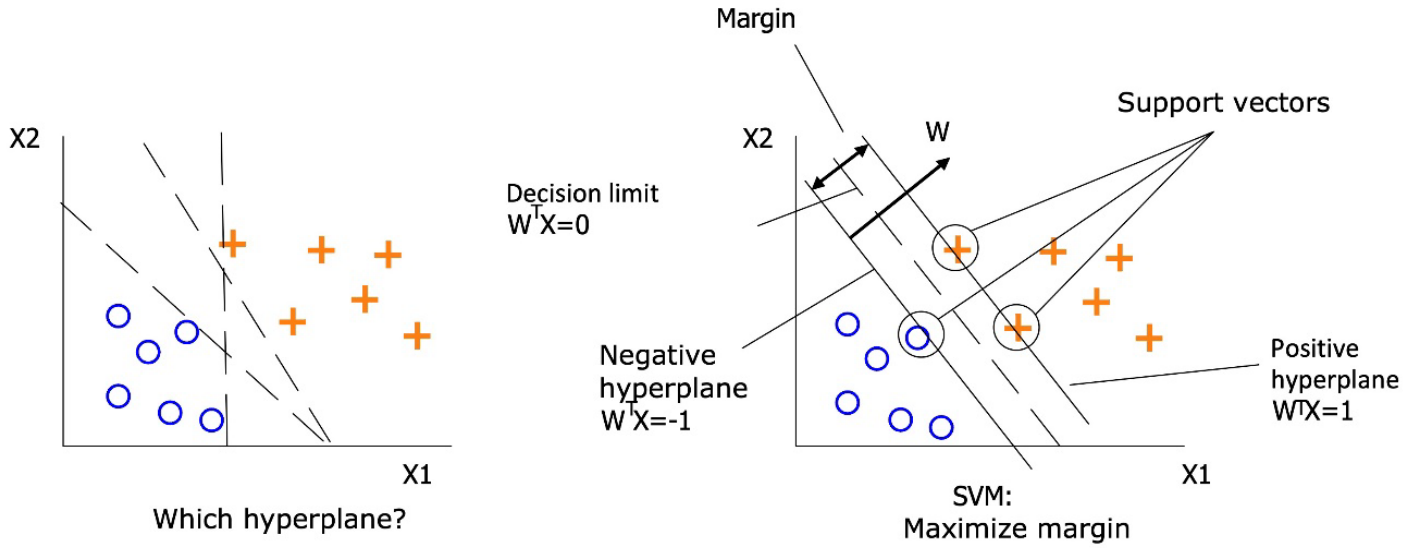

Figure 3. Description of the search for the maximum separation hyperplane between classes in a support vector machine (SVM). Source: Raschka and Mirjalili (2017).

Vapnik (1995) introduced the concept of looseness (ξ, ξ^*) to soften the linear restrictions and converge in optimization of problems that are not linearly separable. The loss function consists of minimizing the L function (Equation (3)) subject to the restrictions (Equations (4), (5), and

(6)). The hyperparameter C controls the width of the margin of separation between classes (Smola & Schölkopf, 2004).

$$\mathrm{L} = \min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi^*{}_i\right)\right) \quad (3)$$

$$y_i - \langle w, \mathrm{x}_i\rangle - b \leq \varepsilon + \xi_i \quad (4)$$

$$\langle w, \mathrm{x}_i\rangle + b - y_i \leq \varepsilon + \xi^*{}_i \quad (5)$$

$$\varepsilon_i, \xi^*{}_i \geq 0 \quad (6)$$

where $\{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_l, y_l)\}$ is the set of training data contained in space X; $\langle w, \mathrm{x}_i\rangle$ denotes the product point between the vector of weights w and the inputs x_i in X; and b is a variable that represents the bias.

The objective is to obtain a function that has a maximum deviation ε of the target classes y_i obtained for all the training data and is as flat as possible, flat means the search for the smallest w. To guarantee this condition, the minimum of the norm $\|w\|^2 = \langle w, w\rangle$ is obtained; this is transformed into a problem of convex optimization (Equation (3)).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Classifier training

To train the classifiers RF, GB, and SVM, a k-fold cross-validation (CV) was performed; this consists of subdividing the original data set into k different subsets. For $k = 10$, ten combinations of training and test data sets are generated in a proportion of 9:1 (MathWorks Inc., 2021). With these subsets, the predictive capacity of the classifiers is evaluated. This technique permits avoiding problems of overfitting the algorithms to the training data. Likewise, stratification is carried out to obtain the same proportions of classes in each selected training and test subset. When training is finalized, k performance values (PG) of the model and the average PG are obtained (Raschka, 2018).

Selection of classifier hyperparameters

Optimal selection of the hyperparameters (fitting parameters that do not depend on the input data) of each classifier was carried out in two stages. The first was selection on a validation graph that describes the variation

of the classification error in function of the values of each selected hyperparameter. In the second stage, selection is fine-tuned through a grid search that consists in defining a search range for each hyperparameter (Table 2). This allows selecting the combination of hyperparameter values that obtains the best performance of the algorithm.

Table 2. Hyperparameter search ranges for the random forest (RF), gradient boosting (GB), and support vector machine (SVM) models.

Model	Hyperparameter	Range
RF	Number of estimators (N)	(1, 500)
	Maximum depth (MP)	(1, 46)
	Minimum samples per node (MMD)	(2, 10)
	Minimum samples per leaf (MMH)	(1, 10)
GB	Learning rate (TA)	(0.0001, 10)
	Number of estimators (N)	(1, 800)
	Minimum samples per leaf (MMH)	(1, 10)
	Maximum depth (MP)	(1, 15)
SVM	*Kernel* (K)	*rbf*, *poly*, *s*, *l*
	Parameter of penalization (C)	(0.1, 100)
	Gamma (G)	(0.1, 1)

rbf: Gaussian, *poly*: polynomial, *s*: sigmoid, and *l*: linear.

The RF and GB ensemble models are based on the standard algorithm of decision trees, whose structure can be controlled by the definition of hyperparameters: number of trees or estimators (N), minimum number of samples to be considered a node (MMD), minimum number of samples to be considered a leaf (MMH), and maximum permitted depth limit (MP), that is, no additional division of this value is allowed (Swamynathan, 2017). The learning rate (TA) refers to the magnitude of the updates of the GB algorithm; a low value delays convergence time with a high probability of selecting a local minimum value, and a high value increases the possibility of skipping the global minimum solution. Selection of this hyperparameter is experimental and with previous knowledge of similar problems (Raschka & Mirjalili, 2017). In SVM, C is the hyperparameter that balances between minimizing the looseness variables and maximizing the margin. Moreover, SVM implements functions of mapping or kernels to augment the dimension of the problem and find a separation plane, which without this transformation would not be possible to visualize. The most frequent kernel is Gaussian (*rbf*, radial basis function) with its *Gamma* hyperparameter (Patle & Chouhan, 2013).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Metrics of performance assessment

To assess the performance of the RF, GB and SVM classifiers, the following metrics were used: global precision of the correct classification (PG), Precision (P), Sensitivity (S), F1_score ($F1s$) and the AUC area under the ROC (receiver operating characteristic curve). These metrics are obtained from a confusion matrix that describes the results of the classifier in a double input table, which shows the predicted values (columns) and observed values (rows) of each class (Figure 4). The confusion matrix allows visualization of the classifier's hits and misses (Barrero-Ortiz, 2019).

Tecnología y Ciencias del Agua
OPEN ACCESS
Check for updates

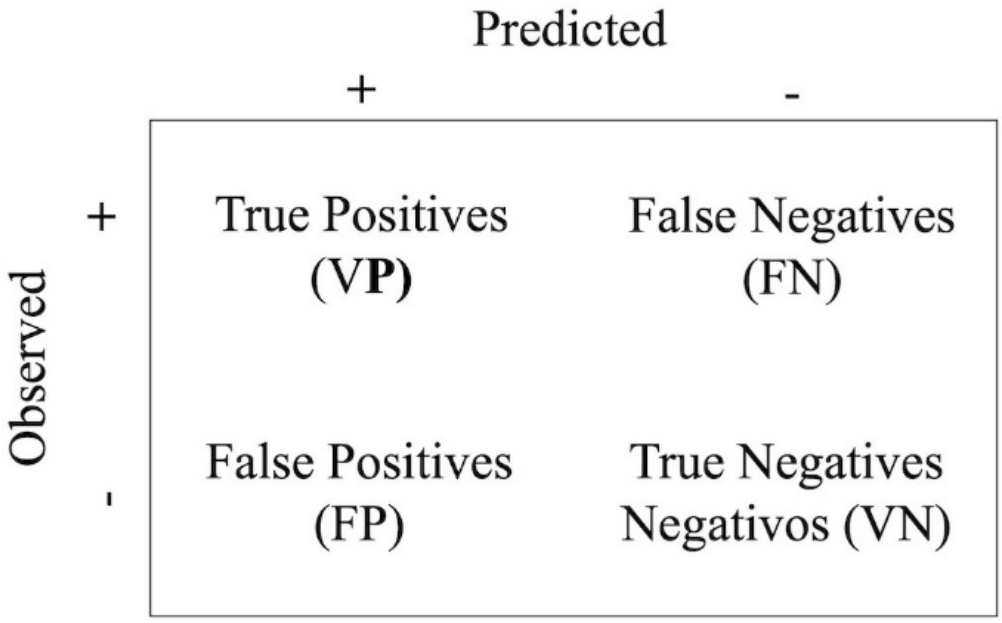

Figure 4. Description of a confusion matrix for the case of binary classification.

$PG = 1 - Ec$ measures the overall proportion of correct classifications, Ec is the classification error; PG is calculated by:

$$PG = \frac{VP+VN}{VP+VN+FP+FN} \quad (7)$$

Where VP is the number of positive samples that are correctly classified; VN is the number of negative samples that are correctly classified; FP is the number of negative samples that are incorrectly

classified as positive; and FN is the number of positive samples that are incorrectly classified as negative; S is the proportion of cases that are correctly predicted as positive, relative to the total observed positive values and is defined as:

$$S = \frac{VP}{VP+FN} \tag{8}$$

P is the proportion of VP relative to the total predicted positive values and is calculated as (Bradley, 1997):

$$P = \frac{VP}{VP+FP} \tag{9}$$

$F1s$ is the harmonic mean of S and P and is expressed as:

$$F1s = \frac{2*S*P}{S+P} \tag{10}$$

where $F1s$ varies in the interval (0, 1).

The ROC measures the discriminating capacity of the model to differentiate the classes and compares the discriminating capacity of two or more models. In a typical ROC graph, each point of the curve corresponds to a possible cutting point of the model and shows its

respective S values (Y axis) and $TFP = FP/(FP + VN)$ (rate of false positives on the X axis). AUC measures how well the model discriminates samples that belong to a class A from others that do not, along the entire interval of possible cutting points (Parikh, Mathai, Parikh, Chandra-Sekhar, & Thomas, 2008; Fawcett, 2006).

Results and discussion

Radar image processing

Figure 5 shows the results of processing the radar image of the study area. Radiometric calibration for the VH bands (Figure 5A) and VV (Figure 5C) standardizes the values of retro-dispersion. *Lee* filters were used to correct the salt and pepper noise as well as later It was applied a geometric correction to WGS84 UTM zone 15 north of the VH bands (Figure 5B) and VV (Figure 5D).

Tecnología y Ciencias del Agua
OPEN ACCESS
Check for updates

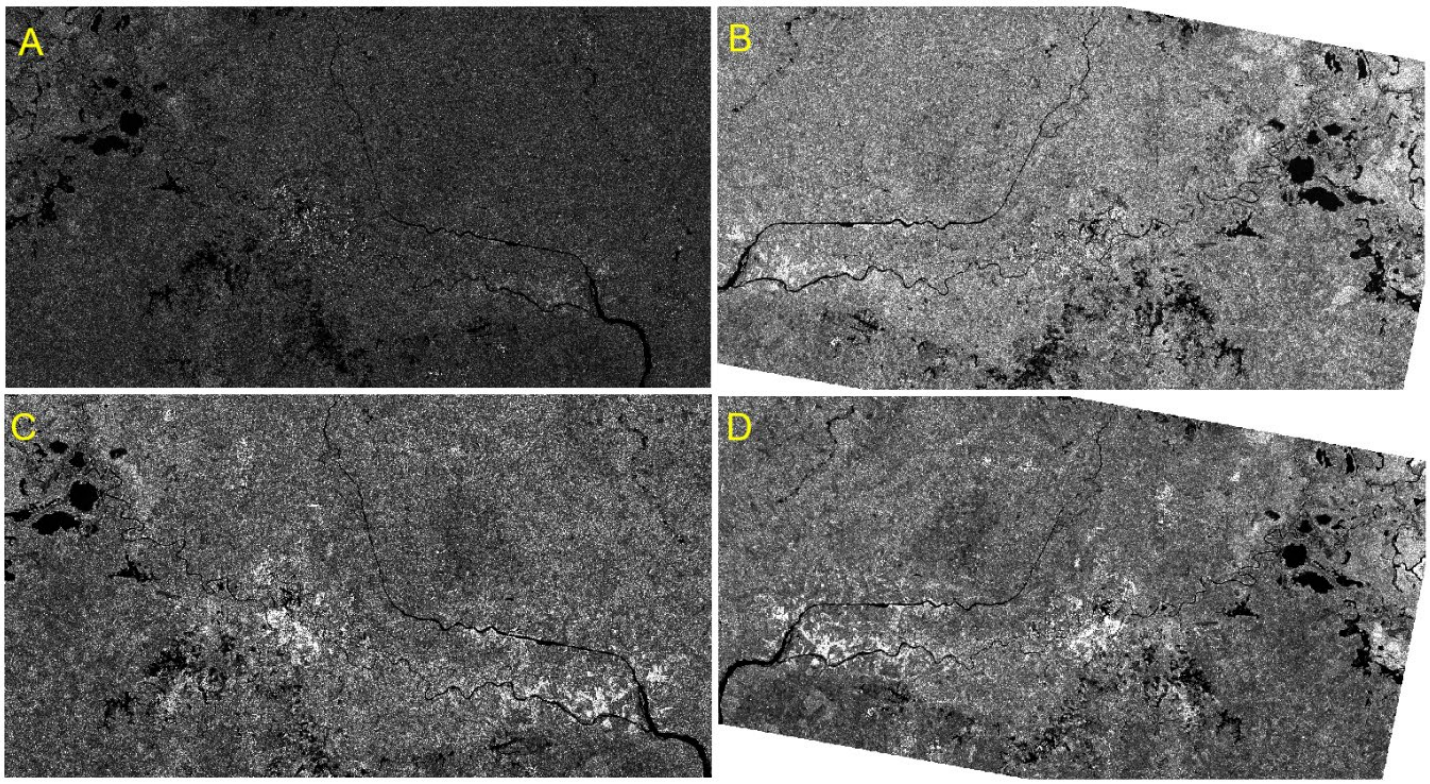

Figure 5. Radar image processing, VH polarization: A) radiometric calibration, B) mottle correction and geometric calibration; VV polarization: C) radiometric calibration and D) mottle correction and geometric calibration.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Selected and labeled samples

As the result of processing and analyzing the SAR images, we obtained an integral database with 12 000 samples, 4000 in each target class (A, I, and V), purged of repeated and inconsistent samples

Hyperparameters of the classifiers

Graphic analysis

The optimal hyperparameters of the RF, GB and SVM models selected in the first stage based on the graphic analysis of classification error curves with cross-validation (k = 10) are described in Table 3.

Table 3. Optimal values of the hyperparameters in the first stage of the random forest (RF), gradient boosting (GB) and support vector machine (SVM) models; average classification error (Ec) and standard deviation (Std).

Model	Hyperparameter	Optimum	Ec	Std
RF	N	200	0.024	0.004
	MP	9	0.022	0.003
	MMD	10	0.021	0.003
	MMH	1	0.021	0.003
GB	TA	0.08	0.022	0.003
	N	200	0.022	0.003
	MMH	1	0.022	0.003
	MP	1	0.022	0.004
SVM	K	rbf	0.027	0.004
	C	120	0.027	0.004
	G	1	0.026	0.005

Figure 6 describes the behavior of the hyperparameters N, MP, MMD and MMH of the RF classifier. The individual effect of each hyperparameter on the average classification error (Ec). The analysis of the Ec graphs

allows visualization of the performance behavior and selection of the hyperparameter with the minimum Ec value (Figure 6).

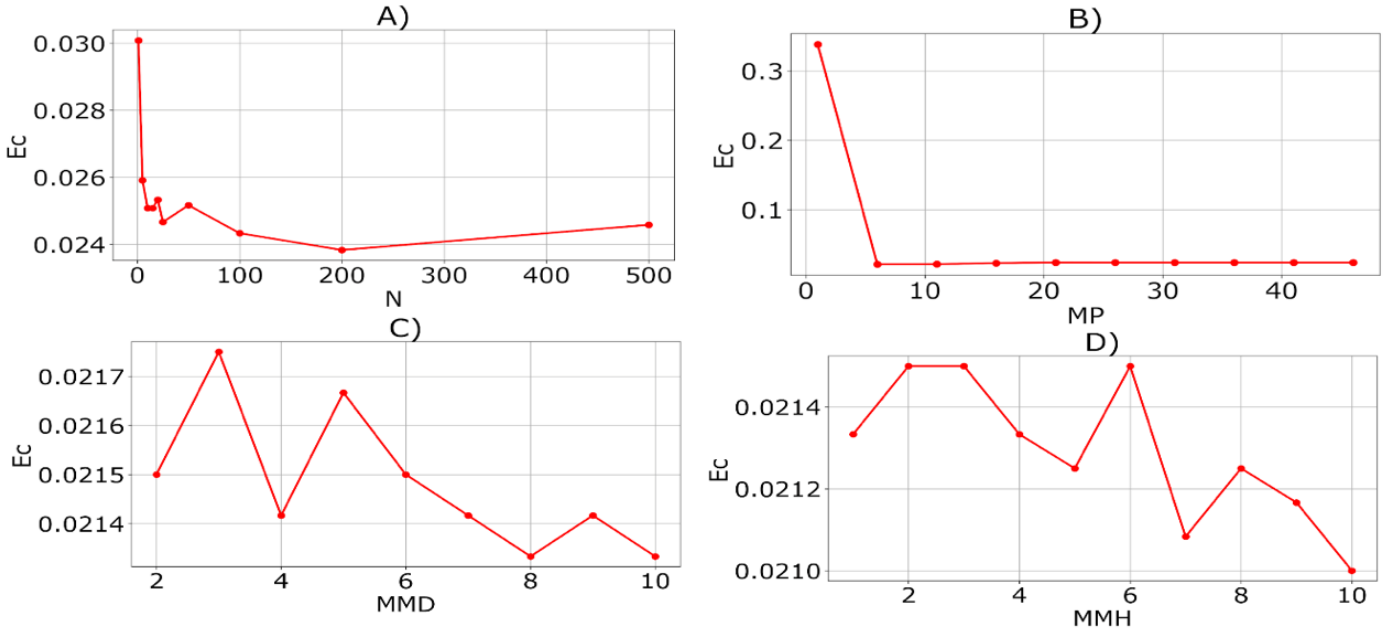

Figure 6. Behavior of the random forest (RF) model classification error in function of the hyperparameters A) N, B) MP, C) MMD, and D) MMH.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Grid Search

To fine-tune the search for optimal values of the hyperparameters of each model, a grid search was carried out with cross-validation (k=10), and the best combinations with the smallest classification error were selected.

The RF model with values $N = 200$, $MP = 9$, $MMD = 10$ and $MMH = 1$ obtained a $PG = 0.979\ (+/- 0.003)$. The GB model with values $TA = 0.08$, $N = 200$, $MMH = 1$ and $MP = 1$ obtained a $PG = 0.979\ (+/- 0.003)$, while the SVM model with values $K = rbf$, $C = 120$ and $G = 1$ obtained a $PG = 0.974\ (+/- 0.005)$.

Based on the grid search, the optimal values of the hyperparameters were different from the values selected with the Ec graph analysis and improved PG performance of the three classifiers. These combinations of optimal hyperparameters were used in the evaluation, comparison and final prediction of the RF, GB and SVM classifiers.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Performance assessment

The assessment metrics of classifier performance obtained by cross-validation are presented in Table 4. The performance evaluation was carried out with a test set of 1200 samples (400 per class). The three classifiers performed better in identification of bodies of water (class A); RF and GB outperformed SVM slightly with $F1s$ of 99.3 %. In contrast, AUC = 1 for the three classifiers means that they are excellent classifiers with adequate selection of the balance threshold between S and P. The high performance achieved for class A is due to its separability from the rest, with values of retro dispersion defined in the interval (-24.0, -22.0) (Fernández-Ordoñez *et al.*, 2020).

Table 4. Performance metric assessment by target class, of classifiers random forest (RF), gradient boosting (GB), and support vector machine (SVM) based on the test set (10 % of the data), with *k* = 10 cross-validation).

Model	S	P	$F1s$	AUC
Class A				
RF	0.998	0.988	0.993	1.00
GB	0.995	0.990	0.993	1.00
SVM	0.995	0.988	0.991	1.00
Class I				
RF	0.963	0.980	0.971	1.00
GB	0.965	0.977	0.971	1.00
SVM	0.950	0.987	0.968	1.00
Class V				
RF	0.980	0.973	0.976	1.00
GB	0.980	0.973	0.976	0.99
SVM	0.990	0.961	0.975	1.000

In identification of physical infrastructure (class I), RF and GB performed better than SVM, with $F1s$ = 97.1 % *versus* 96.8 %. Regarding identification of vegetation (class V), RF and GB performed better than

SVM, with $F1s$ = 97.6 % *versus* 97.5 %. In terms of AUC, the three classifiers performed equally. The performance of the analyzed algorithms decreases due to the confusion between class I and class V, caused by the similarity of the retro dispersion values, translating into false negative predictions for both classes (Figure 7).

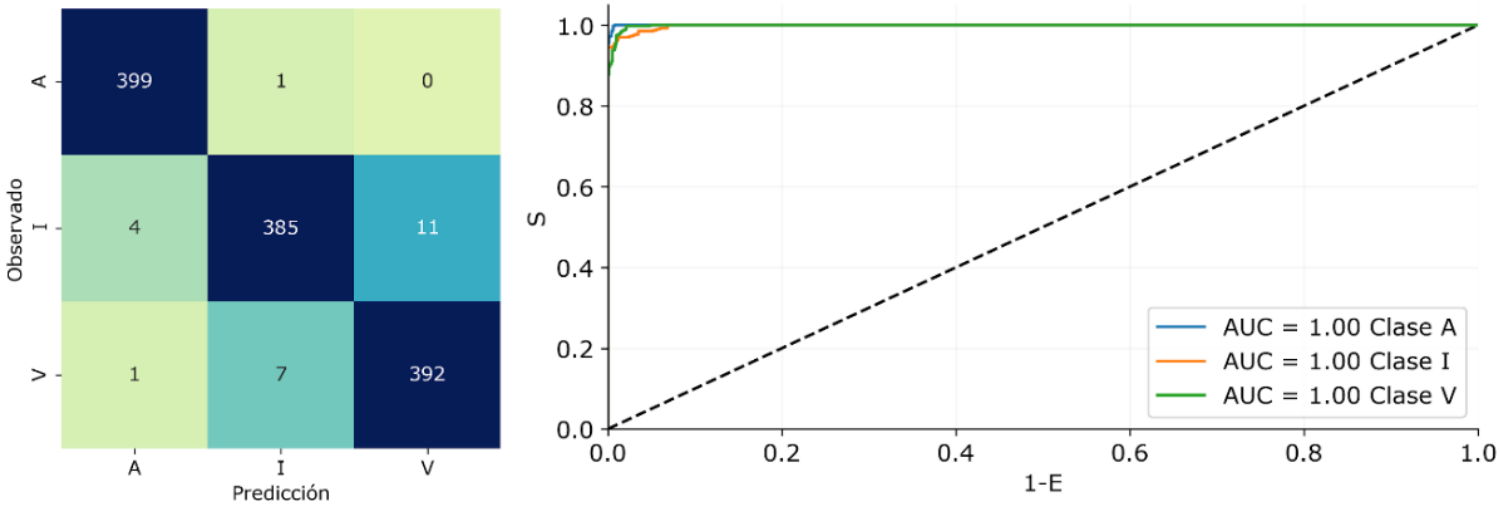

Figure 7. Confusion matrix and ROC of prediction performance of the random forest (RF) classifier based on the test set (10 % of the data), with *k* = 10 cross-validation.

To obtain the best estimator PG, a $k = 10$ cross-validation was carried out. Figure 8 describes the PG behavior of the classifiers RF, GB and SVM. Average PG is represented by a black dot inside each box. RF achieved PG = 0.979 (+/-0.003), GB obtained PG = 0.979 (+/-0.003),

and SVM attained PG = 0.974(+/-0.005). These results show that average performance of the three classifiers is very similar to that obtained with a single test set. Likewise, the ensemble classifiers RF and GB performed slightly better than SVM. For practical purposes, the three classifiers achieved high overall classification accuracy in prediction.

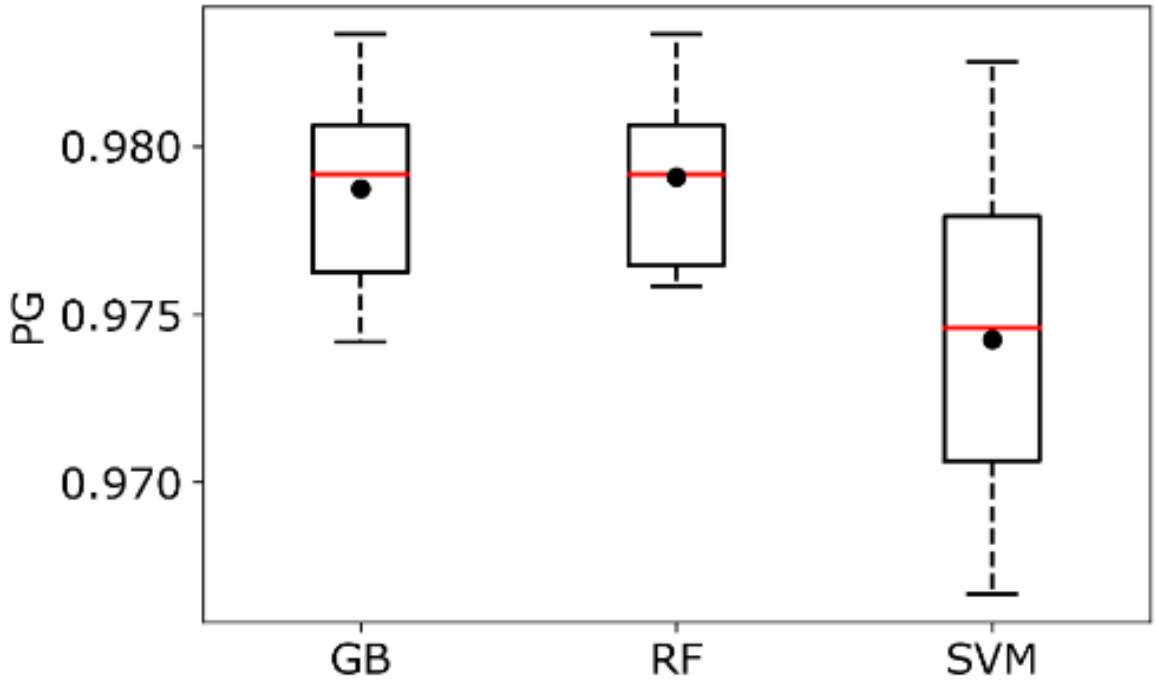

Figure 8. Box plot comparing overall accuracy (PG) obtained by cross-validation (k = 10) of the classifiers gradient boosting (GB), random forest (RF), and support vector machine (SVM).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Study area zoning

The input samples were classified with the trained RF, GB and SVM classifiers and their optimal hyperparameters. The input and output TIF images are georeferenced with the World Geodetic System (WGS84, World Geodetic System 1984) projected in UTM zone 15 north. Figure 9 illustrates the classification of a subsection of the study area that covers 46 842.8 ha (11.5 % of the total area). For each target class A, I, and V, RF classified 3.6, 8.3 and 88.1 %; GB classified 3.7, 8.0, and 88.3 %; and SVM classified 3.7, 15.6, and 80.7 %, respectively. For class A, the percentages of area calculated by the three models were similar. For class I, however, SVM duplicated the estimated area, relative to that estimated by RF and GB. Figure 10 presents the classification of the study area in its entirety (408 687.1 ha). The RF model obtained A = 15 139.2 ha (3.7 %), I = 30 318.8 ha (7.4 %) ha, and V = 363 229.1 ha (88.9 %).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

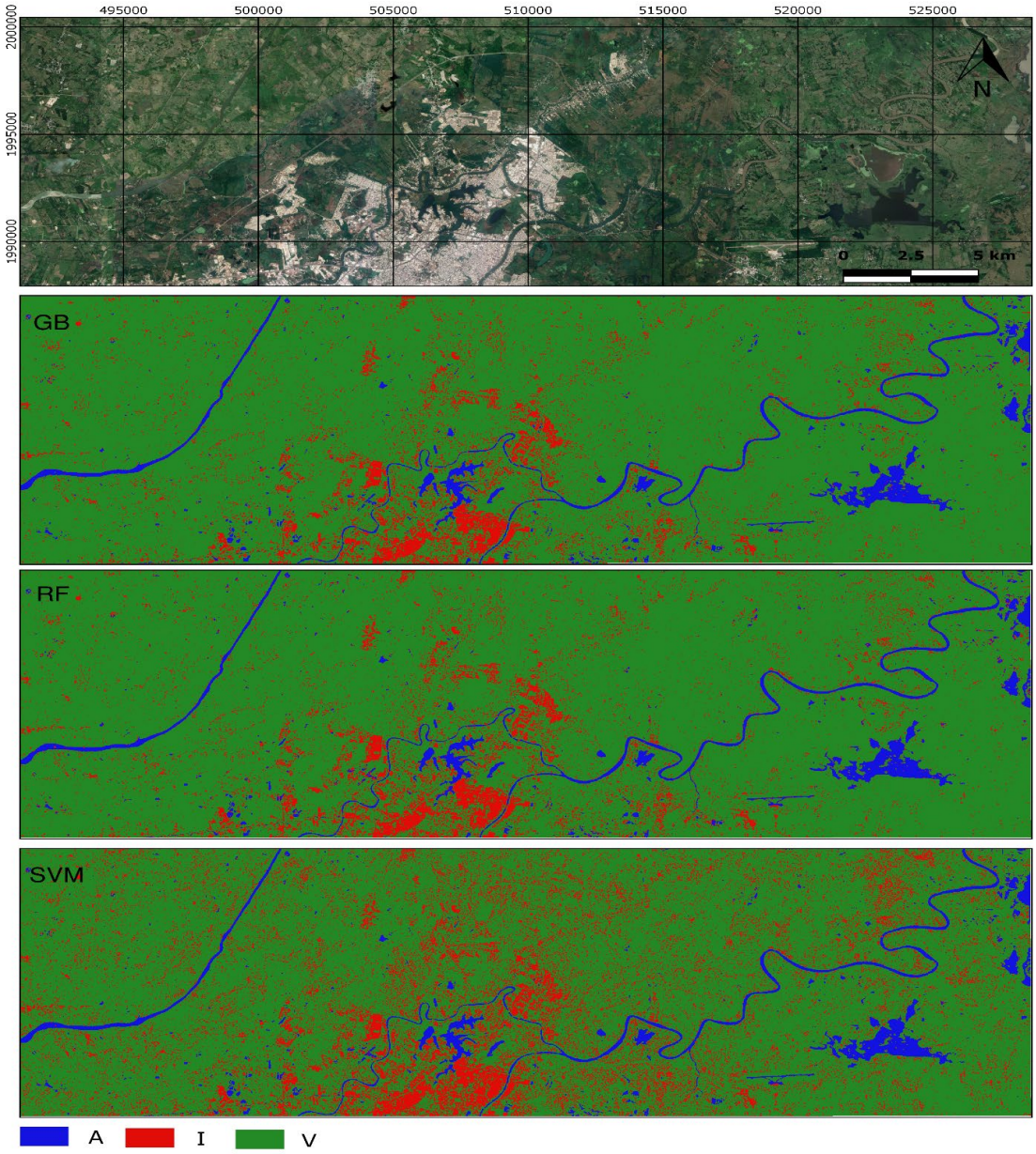

Figure 9. Classification of bodies of water (A), infrastructure and bare soil (I), and vegetation (V) with random forest (RF), gradient boosting (GB), and the support vector machine (SVM) models in a section of the study area located in the states of Tabasco and Chiapas, Mexico.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

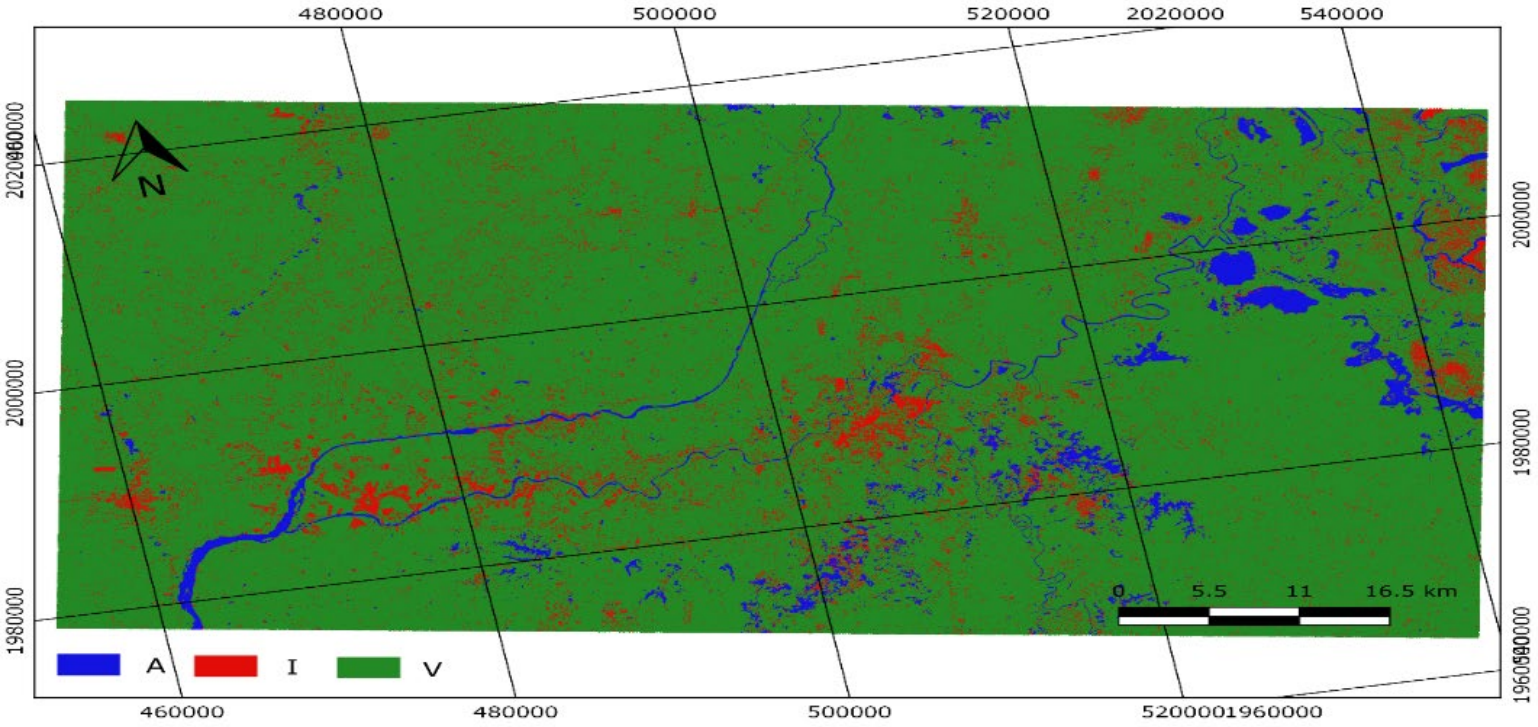

Figure 10. Classification of bodies of water (A), infrastructure and bare soil (I), and vegetation (V) with the random forest (RF) model in the study area between the states of Tabasco and Chiapas, Mexico.

Figure 9 confirms the high performance of the models (RF, GB and SVM) in identifying bodies of water with free area and little aquatic vegetation. The limits of the perennial rivers Grijalva and Usumacinta are perfectly delineated, as are the lagoons with their respective increment in extension. The machine learning models allowed identification of the urban areas of Villahermosa flooded by the intense rainfall on October 6,

7 and 8, 2017, that caused the rivers to overflow. The performance of the classifiers used in this study is slightly better than that obtained by Chen, Huang, Chen and Feng (2021), who used a standard adaptative threshold method to automate identification of flooded areas, obtaining a PG of 95-97 %. In a similar way, the work of Hlaváčová, Kačmařík, Lazecký, Struhár and Rapant (2021),reported PG = 83 % mainly due to the study area and the automated approach that these authors used. The three classifiers correctly identified class I, which corresponded to the urban area of Villahermosa. Nevertheless, they predict a high rate of false positives that correspond to class V. This error is duplicated with the SVM model (Figure 9). The number of identification errors increases when the prediction is realized in the entire study area (Figure 10).

Conclusions

The machine learning classifiers random forest (RF) and gradient boosting (GB) had better prediction performance than the classifier support vector machine (SVM) in identifying bodies of water from synthetic aperture radar (SAR) images. RF and GB had an overall average classification

accuracy (PG) of 97.9 % and SVM PG = 97.4 %. The three models obtained an F1_score above 99.3 % in the prediction of target class A (water); 97.6 % for class V (vegetation), and 97.1 % for class I (infrastructure).

The high prediction performance of the three classifiers is due, among other reasons, to the fact that the target classes (water, vegetation, and soil) were balanced, that the samples of the SAR image associated with each class were separable with the values of the radar bands, and that the grid search of the hyperparameters of each model, based on cross-validation reduced classification errors.

The prediction of areas covered by water with a global accuracy of 99.2 % by the classifiers RF and GB from SAR images shows the potential use of these images in conducting studies related to detection of bodies of water, monitoring, and evaluation of flood damage. Overall, models were highly precise, although they were slightly less precise in distinguishing the infrastructure and vegetation classes.

References

Abdurahman-Bayanudin, A., & Heru-Jatmiko, R. (2016). Orthorectification of Sentinel-1 SAR (Synthetic Aperture Radar) data in some parts of south-eastern Sulawesi using Sentinel-1 Toolbox. 2nd International Conference of Indonesian Society for Remote Sensing (ICOIRS), IOP Conference Series: Earth and Environmental Science, 47(012007). DOI: 10.1088/1755-1315/47/1/012007

ASF, Alaska Satellite Facility. (2020). Distributed active archive center. Recovered from https://asf.alaska.edu/

Arreguín-Cortés, F. I., & Rubio-Gutiérrez, H. (2014). Análisis de las inundaciones en la planicie tabasqueña en el periodo 1995-2010. Tecnología y ciencias del agua, 5(3), 5-32.

Avendaño-Pérez, J., Parra-Plazas, J., & Fredy-Bayona, J. (2014). Segmentación y clasificación de imágenes SAR en zonas de inundación en Colombia, una herramienta computacional para prevención de desastres. Universidad Antonio Nariño, Revista Facultades de Ingeniería, 4(8), 24-38.

Barrero-Ortiz, G. (2019). Evaluación de la eficiencia de los modelos machine learning para la predicción de la calidad del software desarrollado en IBM RPG usando la matriz de confusión y las curvas ROC. Maestría en Gestión de Tecnologías de Información de la escuela de Postgrado de la Universidad Cesar Vallejo de Lima Perú. ResearchGate. DOI: 10.13140/RG.2.2.31760.87040

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159. DOI: 10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Chen, S., Huang, W., Chen, Y., & Feng, M. (2021). An adaptive thresholding approach toward rapid flood coverage extraction from Sentinel-1 SAR imagery. Remote Sensing, 13(23). DOI: https://doi.org/10.3390/rs13234899

ESA & SEOM, European Spatial Agency & Scientific Exploitation of Operational Missions (2019). Sentinel Application Platform (SNAP 7.0). Programa para el análisis y procesamiento de imágenes satelitales (software). Recovered from http://step.esa.int/main/download/snap-download/

Copernicus Sentinel Data (2017). Open access to Sentinel-1 user products. Recovered from https://scihub.copernicus.eu/

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861-874. DOI: 10.1016/j.patrec.2005.10.010

Fernández-Ordoñez, Y. M., Soria-Ruiz, J., Leblon, B., Macedo-Cruz, A., Ramírez-Guzmán, M. E., & Escalona-Maurice, M. (2020). Imágenes de radar para estudios territoriales, caso: inundaciones en Tabasco con el uso de imágenes SAR Sentinel-1A y Radarsat-2. Realidad, Datos y Espacio, Revista Internacional de Estadística y Geografía, 11(1).

Fernández-Ordoñez, Y. M., & Soria-Ruiz, J. (2015). Imágenes de radar de apertura sintética y conceptos básicos de polarimetría. En: Fernández-Ordoñez, Y., Escalona-Maurice, M. J., & Valdez-Lazalde, J. R. (eds.). Avances y perspectivas de geomática con aplicaciones ambientales, agrícolas y urbanas (pp. 37-66). Montecillo, México: Colegio de Postgraduados.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Mach Learn, 63, 3-42. DOI: 10.1007/s10994-006-6226-1

Gini, C. (1912). Variabilità e mutabilità. Reprinted in: Pizetti, E., & Salvemini, T (eds.). Memorie di metodologica statistica. Rome, Italy: Libreria Eredi Virgilio Veschi.

Gomarasca, M. A., Tornato, A., Spizzichino, D., Valentini, E., Taramelli, A., Satalino, G., Vincini, M., Boschetti, M., Colombo, R., Rossi, L., Borgogno-Mondino, E., Perotti, L., Alberto, W., & Villa, F. (2019). Sentinel for applications in agriculture. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-3/W6, 91-98.

Hlaváčová, I., Kačmařík, M., Lazecký, M., Struhár, J., & Rapant, P. (2021). Automatic open water flood detection from Sentinel-1 multi-temporal imagery. Water, 13(23), 3392. DOI: https://doi.org/10.3390/w13233392

Lin, Y. N., Yun, S., Bhardwaj, A., & Hill, E. M. (2019). Urban flood detection with Sentinel-1 multi-temporal Synthetic Aperture Radar (SAR) observations in a Bayesian framework: A case study for hurricane Matthew. Remote Sensing, 11(15). DOI: 10.3390/rs11151778

MathWorks Inc. (2021). Documentación: cvpartition. Recovered from https://la.mathworks.com/help/stats/cvpartition.html

MathWorks Inc. (2016). MATrix LABoratory (MATLAB R2016a). Plataforma matemática sumamente potente en la manipulación de matrices y análisis numérico (software). Natick, USA: MathWorks Inc.

Parikh, R., Mathai, A., Parikh, S., Chandra-Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology, 56(1), 45-50. DOI: 10.4103/0301-4738.37595

Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification. 2013 International Conference on Advances in Technology and Engineering (ICATE), 2013, 1-9. DOI: 10.1109/ICAdTE.2013.6524743

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Perevochtchikova, M., & Lezama-de-la-Torre, J. L. (2010). Causas de un desastre: inundaciones del 2007 en Tabasco, México. Journal of Latin American Geography, 9(2), 73-98.

Podest, E. (16 de abril, 2018). Imágenes Radar (SAR) Seminario NASA - Preprocesamiento, clasificación agua, tierra, Sentinel 1 (2/4) (video). Recovered from https://youtu.be/pVeh9ChATwA

Pulella, A., Aragão-Santos, R., Sica, F., Posovszky, P., & Rizzoli, P. (2020). Multi-temporal Sentinel-1 backscatter and coherence for rainforest mapping. Remote Sensing, 12(847). DOI: 10.3390/rs12050847

QGIS.org. (2020). QuantumGis (QGIS versión 3.10.7-A Coruña). Sistema de Información Geográfica, proyecto de fundación geoespacial de código abierto (software). Recovered from https://qgis.org/es/site/

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. Madison, USA: University of Wisconsin.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow (2nd ed.). Birmingham, UK: Packt Publishing Ltd.

Sami, A., & Abdulmunem, M. E. (2020). Synthetic aperture radar image classification: A survey. Iraqi Journal of Science, 61(5), 1223-1232. DOI: 10.24996/ijs.2020.61.5.29.

Sánchez, A. J., Salcedo, M. A., Florido, R., & Mendoza, J. D. (2015). Ciclos de inundación y conservación de servicios ambientales en la cuenca baja de los ríos Grijalva-Usumacinta. Contactos, 97, 5-14.

Shen, X., Wang, D., Mao, K., Anagnostou, E., & Hong, Y. (2019). Inundation extent mapping by synthetic aperture radar: A review. Remote Sensing, 11(879). DOI: 10.3390/rs11070879

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14, 199-222.

Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps. Apress. DOI: 10.1007/978-1-4842-2866-1

UN-SPIDER, United Nations Platform for Space-based Information for Disaster Management and Emergency Response. (2020). Step-by-step: Mudslides and associated flood detection using Sentinel-1 data. Recovered from https://un-spider.org/advisory-support/recommended-practices/mudslides-flood-sentinel-1/step-by-step

Vapnik, V. (1995). The nature of statistical learning theory. New York, USA: Springer.

Zhang, B., Wdowinski, S., Oliver-Cabrera, T., Koirala, R., Jo, M. J., & Osmanoglu, B. (2018). Mapping the extent and magnitude of sever flooding induced by hurricane Irma with multi-temporal sentinel-1 SAR and InSAR observations. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-3(3), 2237-2244. DOI: 10.5194/isprs-archives-XLII-3-2237-2018