Detection of *Phoradendron Velutinum* Implementing Genetic Programming in Multispectral Aerial Images in Mexico City

Paola Andrea Mejia-Zuluaga, Leon Felipe Dozal-García, and Juan Carlos Valdiviezo-Navarro

Abstract This research implements Genetic Programming to design a spectral index that allows the automated detection of the species *Phoradendron Velutinum* because it is a pest that leads to the detriment of forest health causing serious damage to the host trees. Employing multispectral aerial images taken in the field, pre-processed and selected for the creation of a set of masks with the presence of the pest, together with the use of different terminals and functions, it was possible to obtain an algorithm capable of classifying mistletoe with 96% overall accuracy and a fitness value (Weighted Cohen's Kappa = 0.45) on the test data set. Additionally, a comparison was made with the Structure Intensive Pigment Index 2—SIPI2 for the detection of *P. velutinum*, the results show that SIPI2 does not allow the correct identification of this particular pest.

1 Introduction

Phoradendron velutinum, also known as true mistletoe, is a hemiparasitic shrub plant of aerial parts of trees and shrubs, which has a preference for hosts in broadleaf forests and sometimes with mixed conifers. The spread of this pest has become a problem because it uses a ballistic propulsion method and some animals transport the seed to new host trees, making it difficult to detect new shoots.

P. A. Mejia-Zuluaga (✉) · L. F. Dozal-García · J. C. Valdiviezo-Navarro
Geospatial Information Science Research Center - CentroGeo, Mexico City, Mexico
e-mail: pmejia@centrogeo.edu.mx

L. F. Dozal-García
e-mail: ldozal@centrogeo.edu.mx

J. C. Valdiviezo-Navarro
e-mail: jvaldiviezo@centrogeo.edu.mx

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

R. Tapia-McClung et al. (eds.), *Advances in Geospatial Data Science*, Lecture Notes in Geoinformation and Cartography, https://doi.org/10.1007/978-3-030-98096-2_9

One of the regions affected by this pest is the conservation soil associated with the *San Bartolo Ameyalco* Community located south of Mexico City, where, the uncontrollable presence of the true mistletoe has generated an environmental problem because by reducing the vitality of the trees, and ecological imbalance is presented which leads to a decrease in carbon sequestration and puts the balance of aquifer recharge at risk. and springs that depend on this forest area. Consequently, this region is socially and economically impacted because it is the conservation land that contributes the greatest amount of water to the Valley of Mexico, in addition to reducing the contribution to the Environmental Services administered by the community.

Therefore, this research pursues the objective of designing a spectral index for the automated detection of *Phoradendron velutinum*, generated from the genetic programming and multispectral images, to facilitate early detection and promote adequate control.

Genetic Programming—GP as a tool inspired by evolutionary algorithms, developed to automate tasks and solve different user-defined problems, has been implemented in areas such as regression analysis (Chen et al. 2005), prediction (Huan-rong et al. 2010; Huo et al. 2007), classification (Brameier and Banzhaf 2001), symbolic regression (Icke and Bongard 2013), optimization (Yuan et al. 2008), among others. The diversity of applications is possible since the algorithms of GP perform an abstraction of the knowledge represented in mathematical expressions or systems based on rules, in a probabilistic search space inspired by the Darwinian theory of evolution.

Contributions

Among the contributions made in this research, there is a new approach to the study, detections, and classification of forest pests through evolutionary techniques such as Genetic Programming.

On the other hand, it provides a classification spectral index sufficiently precise to detect the presence of the pest *Phoradendron Velutinum* in the Conservation Soil of Mexico City, and given the characteristics and information requirements for the application of this index, it is possible to do it with conventional aerial RGB images, which implies a reduction in costs and acquisition of equipment for forest sanitation entities such as the Natural Resources and Rural Development Commission—CORENA, with whom We had the opportunity to limit and address the study problem due to mistletoe infestation in the region.

Related Work

There are various methodologies to study pest detection, where the use of aerial images obtained from drones have allowed the inclusion of various supervised learning techniques (Lee et al. 2019; Dwivedi et al. 2021; Shankar et al. 2018).

In the case of forest pests, there are studies on the detection of the bark beetle in mixed forests (Minařík et al. 2020) and the spectral detection of *Dendrolimus tabulaeformis* (Zhang et al. 2020). Additionally, there is research regarding forest parasitic plants such as mistletoe, such is the case of the work reported by Sabrina et al. (2020) in which multispectral aerial images were used with a Convolutional

Neural Network; in this work, they place special emphasis on the architecture of the ANN as an approach to these pests, in the same way, the detection of these species with the use of hyperspectral images was done through the classification of the Spectral Angle Mapper—SAM with a value threshold of 5° considering the spectral signature of the parasitic species. On the other hand, Pernar et al. (2007) reported through a study of mistletoe in coniferous forests in Croatia, that it is possible to detect mistletoe species through a near-infrared digital camera and a combination of supervised and unsupervised learning techniques.

In Mexico, a study was carried out (León Bañuelos 2019) for the detection of the dwarf mistletoe *Arceuthobium globosum* present in the Nevado de Toluca, exploring the implementation of RGB images taken by drone, the supervised classification using the algorithm KNN and colorimetry algorithms. This research is relevant because, in addition to verifying the viability of supervised learning in the exploration and detection of parasitic plants, it also determines under a study that for the detections of forest pests in aerial images, it is necessary to have a spatial resolution less than or equal to 10 cm/px, given the physiological composition of these plants and the little contrast that it could present concerning the host trees.

2 Materials

2.1 Study Area

The study area is located in the conservation land south of Mexico City, registered as Community Conservation Area—ACC of the San Bartolo Ameyalco Community located at coordinates 19°20'N and 99°16'W at a height of 2, 420 *masl* in Álvaro Obregón's Delegation (Fig. 2). That ACC was registered in the Community Areas for Ecological Conservation program in 2017, to protect, improve, and conserve the natural resources and environmental services that this area provides to Mexico City.

The area of the ACC in San Bartolo Ameyalco is made up of 244 hectares, most of which are cataloged as forest area, due to the presence of oyamel *(Abies religious)* forests in 23% of the surface and pine *(Pinus spp)* at 38%. There is also the presence of forests of cedar *(Cupressus)* and oak *(Quercus spp)*, on the other hand, 22% of the surface is used for rainfed agriculture.

2.2 Phoradendron Velutinum

Phoradendron Velutinum (DC.) Oliv.—*P. Velutinum* is a forest pest also known as Barbas, bungu, graft, evil eye, mistletoe, or on the stick (Universidad Nacional Autónoma de México 2022), which is part of the *true mistletoe* species that are normally found in temperate zones of Mexico City (Gonzáles Gaona et al. 2017).

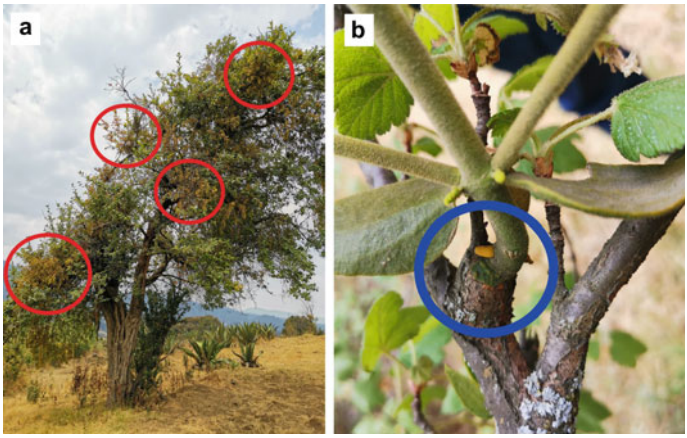

Fig. 1 **a** red circle indicates the presence of *P. Velutinum* at different levels of the host tree. **b** blue circle shows the establishment of the haustorium

P. velutinum is a hemiparasitic shrub with branched stems, opposite and decussate leaves, inflorescences in the form of articulated pedunculated spikes. It generally occurs in the aerial parts of trees or shrubs because the location in the higher parts allows it to obtain sufficient illumination for the photosynthesis process (Fig. 1), and due to its hemiparasitic condition, it is capable of extracting compounds from their host and photosynthesizes organic matter (Gutiérrez Vilchis and Reséndiz-Martínez 1994).

Physiologically, it is up to 80 cm long, internodes 8 cm long, without cataphiles, and with a petiole of 5–20 mm long, the coloration of this parasite is green or yellowish with a velutine pubescence frequently yellow in the young parts and the mature parts glabrescent. It has male inflorescences of 2.5 cm long, 2 to 5 segments of 6 to 40 flowers each in 6 longitudinal rows, the female inflorescence is 4 cm long with fruit, 2 to 4 segments of 12 to 30 flowers per segment in 6 rows (Rzedowski and de Rzedowski 2011).

This species has an establishment stage, in which the seed with chlorophyll endosperm produces simple sugars as a source of energy before germination. Afterward, the incubation stage follows, in which a radicle develops that penetrates the host cortex until it reaches the vascular tissues where it develops cortical haustoria (0.8 –12 cm) (Mathiasen et al. 2008). Once established with aerial stems and flowers, a pollination process begins that can take 4 to 6 weeks. Finally, it disperses the seed through ballistic propulsion, that is, projecting the mature fruit around 15 m, which allows said seed to adhere to the branches of a new host; other means of dispersal are through the feces of birds and squirrels that feed on mistletoe (Alvarado-Rosales et al. 2007).

Table 1 P4 Multispectral features

Features	
Max Speed	50 km/h (31 mph) (P-mode)
	58 km/h (36 mph) (A-mode)
Max Flight Time	Approx. 27 minutes.
GSD*	(Aircraft altitude/18.9) cm/pixel
GNSS	GPS + BeiDou + Galileo
Camera	Six 1/2.9" CMOS
	Max Image Size: 1600×1300 (4:3.25)
	Photo Format: JPEG (visible light imaging) + TIFF (multispectral imaging)
Lens	Focal Length: 5.74 mm (35 mm format equivalent: 40 mm), autofocus set at ∞

*Ground Sample Distance (GSD)

2.3 Data Collection

2.3.1 Surface Sampling

The data collection in the field was carried out on April 13, 2021, at 12:30 p.m. with the support of the Commission on Natural Resources and Rural Development—CORENA technical team, who have information on the location of trees infested by *P. Velutinum* in the area.

At a solar elevation of 57.58° and an azimuth of 103.45°, aerial multispectral images of the affected regions were acquired using a P4 Multispectral Unmanned Aerial Vehicle—UAV, used in precision agriculture applications, environmental monitoring, and in-plant inspection and maintenance. This equipment has six 1/2.9" CMOS sensor camera and a 2 MP global shutter with stabilizer on three axes, composed of an RGB sensor (R (450 ± 16nm) G (56 ± 16nm) B (650 ± 16nm)) for the visible spectrum range and five monochrome sensors that include coverage in the Red Edge spectral band—REG (730 ± 16nm) and Near-Infrared—NIR (840 ± 26nm). Table 1 presents the general characteristics of the UAV.

To cover the largest surface in the shortest flight time, the study area was divided into four polygons (see Fig. 2 and Table 2), the altitude range was kept between 80 and 110 m in height to acquire a GSD value lower than 10 cm, which is the spatial resolution suggested by (León-Bañuelos et al. 2020) for the detection of pests at the tree level. Figure 2 shows the distribution of the flight polygons, which cover a total area of 147.25 ha, which were recorded through 2,565 images each made up of 3 individual files in the bands (RGB, REG, and NIR). The full estimated flight time was 1 hour with 40 seconds.

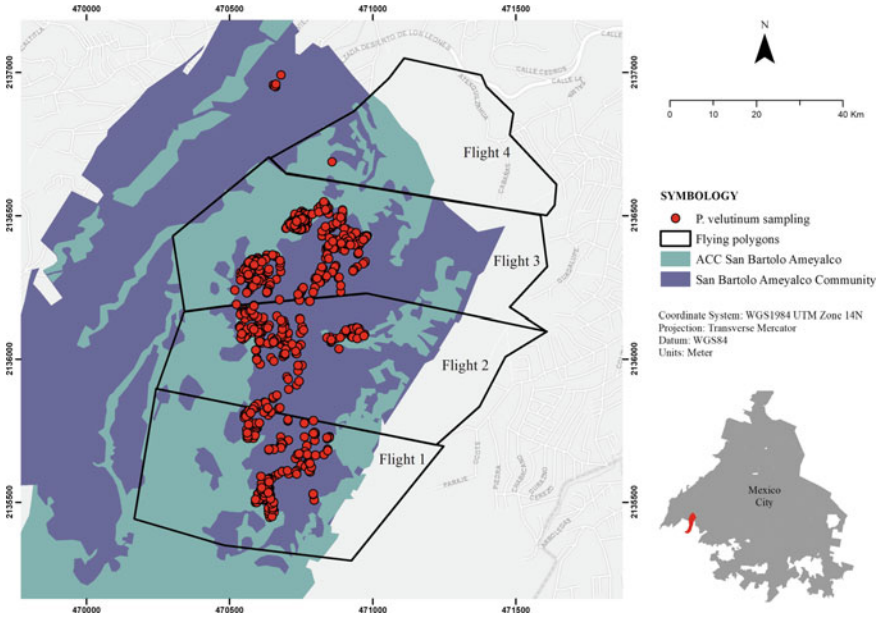

Fig. 2 Flight polygon distribution and sampling of *P. Velutinum*

Table 2 Data collection Design—UAV

Polygon	Area (ha)	Flying height (m)	GSD (cm/px)	Number of images
1	31.40	110	5.8	474
2	46.34	80	4.2	1,244
3	49.14	90	4.8	522
4	20.37	110	5.8	325

2.3.2 Multispectral images

A multispectral image I can be defined as a data cube or pixels $I(x, y, \lambda)$ where x and y represent the coordinates in the plane, column, and row, respectively, and λ represents the depth, that is, the spectral band. In the case of multispectral images, the bands can reach a maximum of 20 (Thigpen and Shah 2008). These images can be obtained by means of filters or sensors sensitive to specific wavelengths, in order to capture different ranges of the electromagnetic spectrum. Usually, ranges of the visible spectrum, are recorded by the combination of the red (R), green (G), blue (B), and infrared bands.

The pixel $I(x1y1, \lambda)$ contains an intensity value called Digital Number (ND) translated from the radiance captured by the sensor (Thigpen and Shah 2008). The ND response can be recorded at different wavelengths that correspond to different bands,

or ranges, into which the Electromagnetic Spectrum is divided cite Emery2017, and depends mainly on the chemical and physical composition of the observed object (Ron Schmitt 2002).

In the case of multispectral images captured with cameras mounted on UAVs such as the *Sequoia* camera, *MicaSense*, or the *P4 Multispectral* sensors, they record the NDs with a spectral resolution of 16 bits (216) which allows storing a greater variety in the grayscale.

2.3.3 Vegetation Index

Given the importance of studying the earth using Remote Sensing techniques, in recent decades the use and applications of spectral information recorded from different sensors such as LiDar, spectroscopes, multispectral cameras, among others, have been deepened. This has led to the creation of spectral indices in the area of Remote Sensing, which, based on mathematical operations, establish relationships between different spectral bands, through which it is possible to detect specific surfaces and patterns. For example, there is a wide range of vegetation spectral indices, which in recent years have been used to detect changes in land cover (Xue et al. 2021), growth models of crops such as cotton (Yeom et al. 2017), or the estimation of the leaf area index (Xie et al. 2014), just to mention a few applications.

The Structure Intensive Pigment Index 2—SIPI2 (Peñuelas and Filella 1995) is one of the indices focused on forest analysis with high foliage variation, this index is related to the structural intensity of the plant's pigmentation (Mihaylov et al. 2020). The computing of this index is carried out using specific wavelengths (800nm, 505nm, 690nm) so it is generally used in satellite images, however, for multispectral images captured with UAV, the implementation is carried out employing the Eq. 1 (Qi et al. 2021; Peres and Cancelliere 2021).

$$SIPI2 = \frac{NIR - Green}{NIR - Red}\ , \tag{1}$$

2.3.4 Computer Equipment

The stage of preprocessing the images to create the data set with the presence of *P. velutinum* (see Sect. 3.2), was performed on a WorkStation Intel (R) Xeon (R) W-2102 CPU @ 2.90GHz, 16 RAM and Windows 10 operating system. The implementation of GP to find an optimal solution to *P detection. velutinum* in aerial images (see Sect. 3.3), was carried out in Software MATLAB R2019a implementing GP-Lab cite Silva2007, on a server Intel (R) Xeon (R) CPU. E5-2690 v2 @ 3.00GHz 250 RAM and Ubuntu 18.04.5 LTS operating system (GNU / Linux 4.15.0-122-generic x86-64.

3 Methods

3.1 Genetic Programming

Genetic Programming—GP is a supervised learning technique established in 1992 by John Koza (Koza 1992) that aims to generate computer programs automatically, is part of Evolutionary Computing, which is a subfield of Computational Intelligence, a branch of Machine Learning and Artificial Intelligence. In Evolutionary Computting, the process of evolution and natural selection described by Charles Darwin in 1859 (Darwin 1859) is described by the survival of the fittest individual where its characteristics are transmitted to the next generation through reproduction and Genetic inheritance is used as a model and as a strategy to find optimal or close to optimal solutions to a given problem. Evolutionary techniques are generally applied to optimization problems. Many of those problems are combinatorial optimization problems, which are computationally difficult (NP-hard).

A GP algorithm works as follows. Initially, a set of possible solutions or candidate solutions called *individuals* is randomly generated, which are evaluated based on an aptitude function that will show how good, or bad, each of these individuals is at the time of solving the problem. Once the aptitude of each individual is evaluated, is used to establish the probability of selection of reproduction of the individuals, that is, of being *Parents* of the next generation. The fitter the individual, the greater their probability of reproduction, this allows the best characteristics to be inherited to the following generations. In this process, the new populations will no longer be random but will be generated through a reproduction process using different *genetic operators* such as crossing and mutation, resulting in new individuals that will also be iteratively evaluated through several generations. until the best possible solutions are found (Koza 1994). Figure 3 shows the general workflow of the GP process.

Representation. An important characteristic of GP algorithms is that they are represented by a tree structure, called *genotype*, which allows increasing computational efficiency compared to their predecessors, Genetic Algorithms (Cramer 1985), since they are determined recursively. Within the tree structure, the root node and the internal nodes contain the functions assigned to solve the problem (operator), and the leaf nodes refer to the data that will be calculated (operand) (Bi et al. 2021), for this example, we will are represented by the spectral bands of the images.

Terminales and Functions. The *terminals* refer to the data or values that will be processed i.e. image bands, numerical variables, vectors, etc., while *functions* refer to the processes or operations that are performed on the terminals i.e. $\{+, -, *, /, cos, sin\}$, etc. Both the terminals and the functions are used to form the structure of the candidate solutions (trees) and must be oriented to establish a pertinent search space to develop the assigned task.

Initial Population. The initialization is carried out by creating the first population of individuals choosing their terminals and functions randomly. There are several procedures for generating the initial population, it is called a generative method, which can take the following directions: total initialization—*fullinit*, increasing

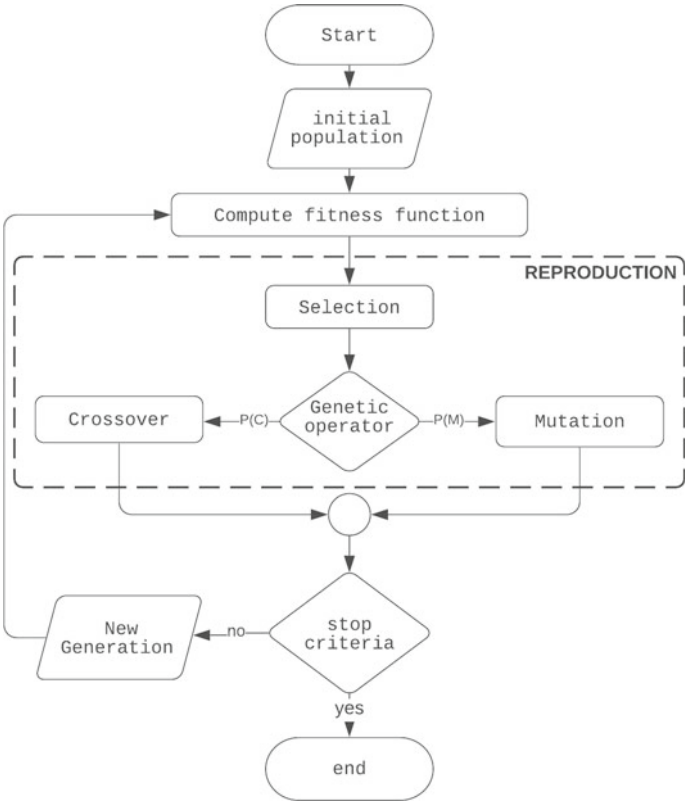

Fig. 3 GP workflow

initialization—*growinit* and ramping initialization—*rampedinit* (Silva and Almeida 2007).

Fitness Function. The calculation of the fitness value of each individual is fundamental for the evolution of the GP, because the "natural" selection, reproduction, and subsequent evolution through the generations, is based on inheriting the best characteristics that lead to perform in an optimal specific task In each generation, the performance of the n individuals of the population must be evaluated to solve the problems, the fitness function must give a certain degree of certainty regarding the problems (Morales and Casas 2007).

Selection. Once the fitness evaluation of each individual has been carried out, it is necessary to select the solutions for their reproduction, where the individuals with the best performance, have a greater probability of reproducing and passing their genetic material to the next generation. However, it is not enough to select the best solutions since when the characteristics of the *Parents* are inherited, the offspring tend to be similar to each other (Poli et al. 2008), which causes a loss of diversity that can be seen represented in a premature convergence.

In the process of selecting individuals, two important genetic parameters are considered (Chaudhary and Iqbal 2009): *sampling* and *elitism*. The sample allows to select the individuals who will act as parents of the next generation, this process can be done through the *roulette* methods, *tournament'*, *lexictour* or *sus*. On the other hand, elitism is how individuals are chosen for the next generation taking into account their evaluation of fitness, among the options is *halfelitism*, *totalelitism*, *keepbest* and *replace*. The technical specifications of each sampling method and elitism can be seen in greater detail in Silva and Almeida (2007).

Genetic Operators. Genetic operators are processes for generating new *children*, from existing *parents*, by combining their trees or substituting branches for new ones. In this study, we use two genetic operators, crossover and mutation. The selection of these genetic operators is assigned employing probabilities, trying to conserve the best genetic material while diversifying the solutions (Banzhaf et al. 1997).

Crossover: consists of the crossing of genetic information provided by two *Parents* to generate one or two *children* from the selection of subtrees (portions of trees assigned from a random node) in each individual, the selected subtrees are exchanged between one individual and another, resulting in offspring with new genetic characteristics.

Mutation: this process aims to introduce new genetic material to reduce bias and increase diversity, through the random selection of a node that will be replaced by new functions and randomly selected terminals (*Grow* method).

3.2 Pre-processing

3.2.1 Image Registration

Before training the GP algorithm, it was necessary to preprocess the images, starting with the registration of the RGB bands with the REG and NIR bands, since a mismatch was experienced between the exposure time of each sensor, causing spatial displacements between the bands of each photographic capture. This step is important for accurate spectral measurement of the presence of *P. velutinum*.

The bands were registered using the algorithm *Speeded-Up Robust Features*—SURF (Bay et al. 2006), this algorithm is designed as a detector and descriptor of local characteristics. It is based on *Scale-invariant feature transform*—SIFT (Lowe 1157), however, it presents improvements regarding the decrease in computational time and more robust support against the transformations on the image. This is because it uses box filters (using integral images (Crow 1984) to create different scale-spaces instead of iteratively reducing the size of the image as in other methods.

The main workflow of SURF consists of interest point detection, local neighborhood description, and point matching.

To carry out the correct registration between bands of the 2, 565 images taken with P4 Multispectral, and since SURF functions are composed of multiple parameters, it was necessary to perform an iteration of these criteria in a sample set of 100

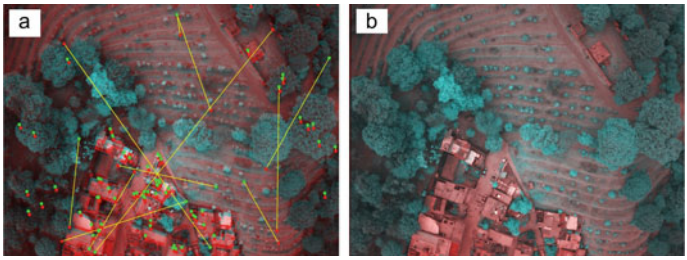

Fig. 4 **a** Points of interest between bands detected from SURF algorithm. **b** Registered Image Result

RGB images together with a random selection of their pairs in both the REG and NIR bands. The computational time to achieve the iteration of 14, 400 combinations was 21 days and 8 hours. The objective of this process was to find the parameters that maximize the registration between the RGB, REG, and NIR bands in a forest environment with a few buildings and roads. On the other hand, the metric used to evaluate the registry was the Structural Similarity Index—SSIM, this index evaluates the similarity between two images without compressing or distorting the reference image; the basis for this metric is a weighted comparison of luminance, contrast, and image structure.

The Fig. 4a, shows an example of the detection of points of interest detected by SURF where you can see the offset between the bands, in part b you can see the results of the successful registration of the images.

3.2.2 Mask Image Creation

Field trips were carried out around various sectors in the study area, to carry out visual training in the detection of *P. Velutinum*, both in its coloration and in its appearance and arrangement on the trees. The knowledge of these aspects in a physical environment as in a digital environment (photographs), is essential for the subsequent identification of the plague on the images.

The identification of the true mistletoe was carried out on the images through image interpretation, based on 885 sampling points GPS captured in field trips carried out by CORENA during the first months of the year 2021. This process resulted in the previous selection of a total of 94 images in which there is a presence of the pest in different proportions. In this selection, the reduction of overlap between exposures was considered to have unique data within the segmentation of the images.

This identification is accompanied by the creation of binary masks using ROI polygons, in which the zones with a value equal to 1 represent the presence of *P. velutinum* and the values equal to 0 represent the absence of it. This creation of masks was carried out through the application *Image Segmenter* (MATLAB 2019) implemented in the Image Processing and Computer Vision toolbox of MATLAB.

Through this application, it was possible to manually locate the sources of mistletoe infestation and segment the polygons present in the image.

3.2.3 Data Set Creation

Once the masks of the 94 images had been made, a division of the original data with a size of 1300 × 160 px was carried out in small pieces of 300 × 300 px. This process is intended to increase the volume and diversity of the data set with different perspectives of the pest on the images and to reduce processing time through the use of smaller data.

Next, it was necessary to make a selection of the samples that first had the presence of mistletoe, since the cut of each image was made in different views sequentially and proportionally to it, many of the samples only refer to the background of the territories (understood as any surface that is not *P. velutinum*). The second requirement implemented in this selection was the balance of classes (Goncalves and Silva 2013), since in some sections the polygon that indicated the presence of the pest represented less than 10% of the total image. manually, considering the details and context of each image.

It is important to highlight that the creation of this data set is based on original information taken in the field to address the specific study of the parasitic plant *Phoradendron velutinum* present in the study area and on the identification, segmentation, and selection by part of the expertise acquired in this work.

3.2.4 Threshold Selection SIPI2

Since Structure Intensive Pigment Index 2 is not a specific index for the classification of mistletoe but rather they present a wide-scale thought in the variety of foliage, it is not possible to determine *a priori* which is the proportion of the index that could best determine the presence of *P. velutinum*. Therefore, a series of threshold tests were carried out to determine which is the most appropriate range for the detection of this pest. This thresholding was performed on 100 randomly selected images within the data set designed for mistletoe (multispectral tiles and their respective mask), the iteration contemplated 10 threshold values ranging from 0 to 0.9.

As a result, it was obtained that the threshold that could best classify the presence of mistletoe is 0.9, a value of $k_w = 0.0047$, $k = 0.0059$, and an $OA = 0.5383$ (Fig. 5).

3.3 GP *Implementation*

The implementation characteristics used to design an optimal rate in the detection of *P. velutinum* by GP are listed below:

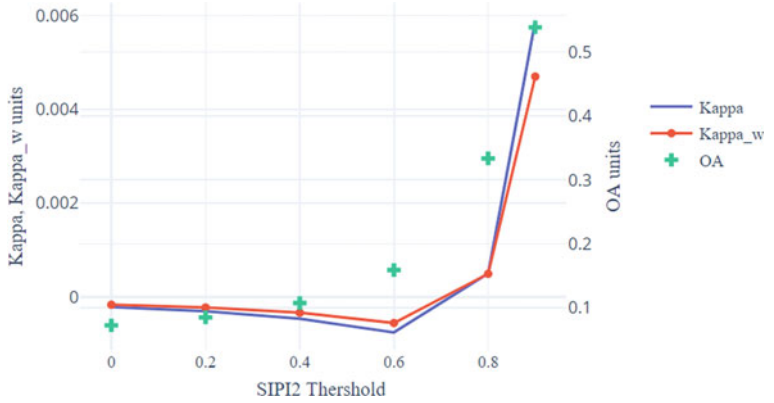

Fig. 5 The thresholding of the SIPI2 values shows how, from thresholds greater than 0.8, minimal changes are possible in the presence of different foliages in some plants, which does not imply that it is *P. velutinum* specifically. OA values reach a range of 0.5 because the fund counts as correctly classified securities

Terminal set. The implemented terminals refer to the bands captured by the multispectral camera: Red (R), Green (G), Blue (B), RedEdge (REG), Near Infrared (NIR).

Set of functions. The implemented functions include the range of arithmetic functions such as addition, subtraction, multiplication, division, and logarithms.

Population initialization. The initial population was generated in a search space under the *rampedinit* Ramped Half-and-Half method, where half of the population is initialized with the *full* method and the other with the *Grow* method, obtaining a population that is half balanced and half unbalanced, reflecting a high level of diversity. In the *full* method, each tree generates internal nodes until it reaches the assigned initial depth, and in the last depth level, it assigns terminal nodes (leaves), the individuals of this method are balanced and with branches of the same size. On the other hand, the *Grow* method randomly integrates a terminal or a function at each node, varying its depth between 1 and the maximum level, resulting in unbalanced trees.

Fitness function. The proposed function to evaluate performance is weighted Kappa (K_w), which is given by Eq. (2). This performance measure is used to evaluate the classification between evaluators, since it is a weighted estimate of the agreement between classes, evaluating for this case n individuals in 2 categories that are exclusive between them (presence or absence of *P. velutinum*).

Since it is a binary classification, a 2×2 confusion matrix is used in which the rows represent the real values (*Ground truth*) while the columns represent the prediction values. The stored values are true positives—*TP*, false positives—*FP*, false negatives—*FN*, and true negatives—*TN* (Banzhaf et al. 1997).

- *TP*—It is a pest and was classified as a pest
- *FP*—Not a pest but classified as a pest

- *FN*—Yes it is a pest but it was classified as not a pest
- *TN*—Not a pest and classified as a non-pest

$$K_w = 1 - \frac{\sum_{i-1}^{k} \sum_{j-1}^{k} w_{ij} x_{ij}}{\sum_{i-1}^{k} \sum_{j-1}^{k} w_{ij} m_{ij}}, \tag{2}$$

where k is the number of categories, w_{ij} are the elements of the weight matrix, x_{ij} refers to the observed matrix and m_{ij} are the values in the matrix expected (Stehman 1996; Cohen 1968).

The weight matrix penalizes erroneous predictions, in this case, the *FP* will be given double weight than the *FN* $\begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$, since it is more expensive in terms of time, financial and human resources to classify an area with mistletoe infestation, mobilize a commission to carry out the appropriate forest health activities and find that there is indeed no presence of the mistletoe. plague.

Selection criteria. For the selection of individuals, the type of sampling selected was the method *lexictour* (Luke and Panait 2022), which consists of the random selection of individuals that will be compared with the rest of the population, from which the best are chosen. In this comparison, only which individual is better is evaluated in parallel with another. However, if two individuals obtain the same performance value, the one chosen will be the one that has the smallest tree, that is, it has fewer nodes in-depth, implying that it will be a simpler solution to implement. On the other hand, the selected elitism method is *Keepbest* giving survival priority to the *Parents* and *children* with the greatest cliff in the weighted Kappa function, while the rest of the population follows an order in which parents come first and children second.

Data set: The proposed data set consists of 250 multispectral images (bands: R, G, B, REG, NIR) of size 300 x 300 px with the presence of *P .velutinum* and 250 segmented images (binary) of size 300 x 300 px with presence or absence of the pest, images that were preprocessed as mentioned in Sect. 3.2 (Fig. 6).

3.3.1 Experimental Design

The objective of the implementation of GP in this investigation is to find a computer program—*a spectral index* that allows the identification of *P. velutinum* on multispectral images. For this, the parameters mentioned in Table 3 are defined, where each experiment was carried out 30 times independently with different training sets. At the end of the 50 proposed generations with *kfold(5)* and implementing *Saliency Toolbox* (Walther 2006), the best individual in each experiment was evaluated with the corresponding test set to evaluate the final performance of the best candidates found. The computational time used in the execution of each iteration was 7 to 8 days for each one.

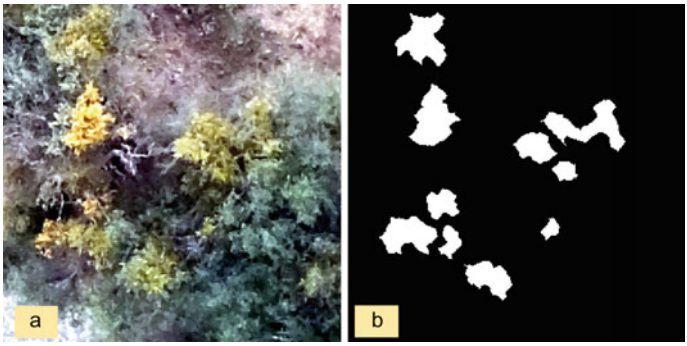

Fig. 6 *P. velutinum* Data set. **a** RGB image with the presence of *P. velutinum*. **b** Binary mask of presence/absence of the pest

Table 3 Parameters used for GP training

Generations	50
Population size	50
Initialization	Rampedinit
Crossover	Probability 0.7
Mutation	Probability 0.3
Selection	Lexictour
Elitism criteria	Keepbest
Functions set	Arithmetic operators
Terminals set	R, G, B, REG, NIR
Fitness function	Weighted Cohen's Kappa

4 Results and Discussion

From the experiments carried out, the best 30 individuals were selected (one for each experiments) and evaluated against the test data set to check their performance in the tasks of detecting the pest *P. velutinum* on multispectral aerial images. For this, the best aptitude values, the least number of nodes, and the diversity of solutions among individuals were taken into account.

Based on the diversity of functions and terminals used in GP training, it is important to discriminate their operating frequency, to study the search space, and analyze the best individuals. Figures 7 and 8 show the frequency of use in the best individuals of the terminals and functions respectively.

The test of the individuals was carried out through the confusion matrix, where additional performance metrics such as (*Overall accuracy*)—OA were considered, indicating the quality of the classification, (*Recall*)-R which reflects what proportion of the relevant elements (*P. velutinum*) that have been detected successfully, Precision—P as the predictive value positive, Cohen's kappa coefficient κ and the

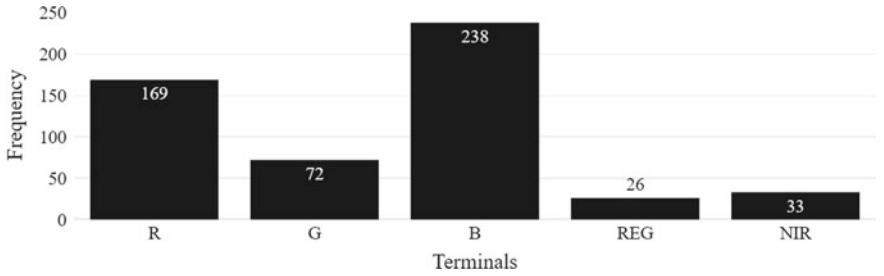

Fig. 7 Terminals frequency. The most used terminals are the Blue band and the Red band

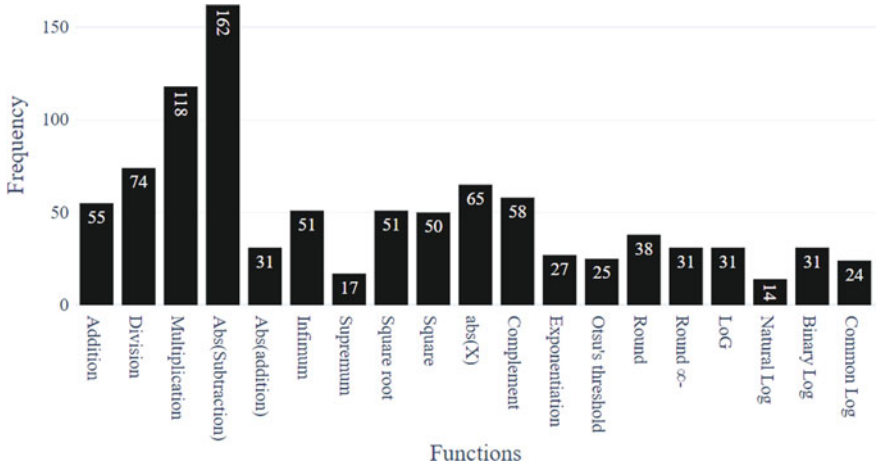

Fig. 8 Functions frequency. The most used functions are the absolute value of subtraction, multiplication, and division

Table 4 Comparison of metrics in the test and training sets—Best GP individuals

Training	Test					Diff.
K_w	K_w	k	OA	R	P	
0.4474	**0.4576**	**0.4311**	**0.9667**	**0.4086**	**0.5661**	0.0102
0.3738	0.4199	0.4082	0.9636	0.4288	0.4860	0.0461
0.3907	0.4137	0.394	0.9625	0.3972	0.5031	0.0230
0.3767	0.4084	0.3998	0.9644	0.4292	0.4737	0.0317
0.3517	0.3991	0.3939	0.9581	0.4518	0.4503	0.0474

fitness function defined for this problem *Weighted Cohen's Kappa*—K_w, the average of the metrics in the test set of the best 5 individuals is shown in the Table 4.

Given the results of the performance metrics of the best GP individuals, the best individual reports a fitness value on the test set of $K_w = 0.4576$ demonstrating greater generalizability to the detection problem of *P. velutinum* on aerial imagery while meeting an overall classification accuracy of approximately 97%.

Table 5 Comparison of metrics in the test and training sets—SIPI2 Index

Training	Test					Diff.
K_w	K_w	k	OA	R	P	
0.0016	−0.0014	0.0984	0.4953	0.5106	0.0356	−0.0030

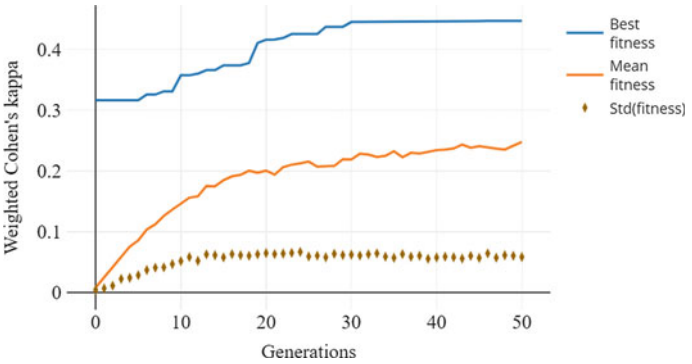

Fig. 9 Evolution of GP's individuals for the optimization of the best fitness and average fitness

On the other hand, the evaluation of the SIPI2 index was carried out on the same set of training and test data used by the best individual obtained from GP (Table 5). We can see that the fitness value obtained by this index is −0.0014 with an overall accuracy of 0.49, this is because when classifying large proportions of the image as mistletoe (because SIPI2 allows us to see changes in the health states of the tree, not necessarily caused by *P. velutinum*) increases the chances of hitting the search space; however, a large percentage of the classification is incorrect as shown by the value of k_w.

Figure 9 shows the evolution of the optimization of individuals through the generations, showing a high rate of convergence in the first 20 generations and a stabilization of the best fitness in generation 30. Average fitness throughout all generations is on average 0.3 units down throughout the evolution with a generalized standard deviation from generation 10. It is important to note that across the evolutionary process it is evident that generationally there are better solutions to the problem of detection of mistletoe in multispectral imaging.

$$
\begin{aligned}
&((abs(\mathbf{B} - (log_{10}(\mathbf{N}))^2) \cdot log_2(abs(\mathbf{B} - \mathbf{R}))) - \mathbf{G}) \cdot \\
&(sup\left(\sqrt{\lfloor \subset (abs(\mathbf{B} - \mathbf{R}) \cdot log_2(\mathbf{N}))\rfloor}, \frac{\lfloor \subset (round(\subset (\mathbf{R}))^2)) \cdot Log_2(\mathbf{N})\rfloor}{\mathbf{B}}\right) \cdot \\
&(abs(\mathbf{B} - \mathbf{R}) \cdot Log_2(\mathbf{G}))) ,
\end{aligned}
\tag{3}
$$

As a result of the test stage, the spatialization of the successes and errors is obtained for both the best GP individual and the SIPI2 index, as shown in Fig. 9, in

the first column the original sample images, in the second the classification performed by the best individual (the best spectral index found as a solution to the detection of *P. velutinum* with GP), the third column shows the SIPI2 index throughout the thresholds and the last column shows the classification performed with SIPI2 for *P. velutinum* at a threshold of 0.8 (previously defined in Sect. 3.2).

The well-classified background can be seen in black, the *P species in white. velutinum* correctly classified, the magenta color represents the commission errors (where mistletoe was detected but there was not) and finally, the omission errors are represented in green (where there is mistletoe but it was not correctly classified).

5 Conclusions

In this research, the implementation of Genetic Programming—GP was carried out for the detection of the parasitic species *Phoradendron Velutinum*, reaching as a whole (considering the entire surface that does not refer to mistletoe) an accuracy greater than 96%. This classification was carried out employing the execution of the best evolved spectral index (individual).

From the results of training, testing, and review of the classification with the best algorithms obtained with GP, it is possible to see that in the generalization of the Phoradendron velutinum detection task on multispectral aerial images, a fitness value of 0.45 is reported but in particular cases, results of up to 0.81 are achieved in Weighted Kappa with a Recall of 0.69 and Precision of 0.89. These results and the visual analysis performed on the best and worst classifications indicate that the assigned weight matrix has a great influence on some misclassified mistletoe shoots, but also indicates that errors of omission can be caused by residuals in the registration process. between the bands of the images, additionally coupled with an error threshold found at the edges of the mistletoe since its extensions (branches and leaves) intertwine with the structure of the host tree, making it impossible to define the limit of the parasitic plant.

On the other hand, it is worth mentioning that SIPI2's approach is to evaluate the different types of chlorophyll in plants to estimate the different types of canopies and foliage, identifying certain deficiencies in the plants. However, since *P. velutinum* is a parasitic plant, it is difficult to discriminate between the vitality and the foliage of the parasite and its host; that is, the SIPI2 index cannot be used for the specific detections of *P. velutinum*.

6 Future Work

The GP technique demonstrates its versatility and adaptation for the study of different territorial problems such as the problem of forest pests addressed in this work. However, it is important to mention that given the high heterogeneity of forest areas

such as the Conservation Land of Mexico City, there are still sections for improvements in the workflow and parameters considered here. It is important to delve into the use of the proposed methodology because despite reaching an average Overall Accuracy of around 96% in the first experiments, several misclassification errors were overlooked in which mistletoe was not correctly detected, generally in areas close to those already correctly classified due to the irregular shape of the parasitic species. On the other hand and more importantly, there are several errors in which it was classified as *P. velutinum* but in reality, this pest did not exist in that area, which is reflected in an average Precision of 56%. It is important to mention that these *commission* errors can present a high cost in forest sanitation activities.

Additionally, it should be noted that it is necessary to deepen the study and analysis of better individuals, for this, it is necessary to continue with a greater number of experiments, which implies a greater investment in processing time.

From the development of this work, several questions of improvements arise both at the level of the acquisition of images, recording of additional spectral bands, the study of the best functions, the best parameters, and the evaluation of performance compared to other algorithms of machine learning classification.

Acknowledgements The authors would like to thank Ing. Israel Jimenez and the Commission for Natural Resources and Rural Development—CORENA for the support and contribution of their experience and accompaniment in the field for the detection of *P. velutinum* in the conservation soil of the City of Mexico.

References

Alvarado-Rosales D, Equihua-Martínez A, López-Gomez Tagle E, Rodríguez-Ortega A, de Lourdes Saavedra-Romero L, Vanegas-Rico JM (2007) Situación actual de la declinación del oyamel en el parque Desierto de los Leones, D.F. In: Memoria del XIV Simposio Nacional de Parasitología Forestal. Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, pp 24–28

Banzhaf W, Keller RE, Francone FD, Morgan MB (1997) Genetic programming: an introduction. Morgan Kaufmann

Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features. In: Leonardis A, Bischof H, Pinz A (eds) Computer vision - ECCV 2006. Springer, Berlin, Heidelberg, pp 404–417

Bi Y, Bing X, Mengjie Z (2021)Genetic programming for image classification. Springer Nature

Brameier M, Banzhaf W (2001) A comparison of linear genetic programming and neural networks in medical data mining. IEEE Trans Evolut Comput 5(1):17–26

Chaudhary UK, Iqbal M (2009) Determination of optimum genetic parameters for symbolic non-linear regression-like problems in genetic programming. In: INMIC 2009 - 2009 IEEE 13th international multitopic conference, vol 042, pp 12–16

Chen J, Li ZZ, Liao ZG, Wang YL (2005) Distributed service performance management based on linear regression and genetic programming. In: 2005 international conference on machine learning and cybernetics. ICMLC 2005, pp 560–563

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bullet 70(4):213–220

Cramer NL (1985) A representation for the adaptive generation of simple sequential programs. In: International conference on genetic algorithms and the applications, pp 183–187

Crow FC (1984) Summed-area tables for texture mapping. Comput Gr (ACM) 18(3):207–212

Darwin C (1859) On the origin of species by natural selection
Dwivedi R, Dey S, Chakraborty C, Tiwari S (2021) Grape disease detection network based on multi-task learning and attention features. IEEE Sens J 21(16):17573–17580
Goncalves I, Silva S (2013) Balancing learning and overfitting in genetic programming with interleaved sampling of training data. In: Krawiec K, Moraglio A, Hu T, Etaner-Uyar AŞ Hu B (eds) Genetic programming. EuroGP, vol 7831. LNCS (December 2016)
Gonzáles Gaona E, Serrano Gómez C, De Lira Ramos KV, Quiñonez Barraza S, Sánchez Martínez G, López Pérez I, Sánchez Lucio R (2017) Identificación, distribución y control de muérdago enano (Arceuthobium spp.) en bosques de coníferas. Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Aguascalientes
Gutiérrez Vilchis LH, Reséndiz-Martínez JF (1994) Fenología del muérdago enano en el desierto de los leones, D.F. Revista de Ciencia Forestal en México 19(75):41–62
Huan-rong Z, Ya-min L, Ia-mei M (2010) Based on meteorological factors and short-term load forecasting genetic programming. In: 2010 International conference on computer design and applications, vol 3. IEEE, pp 465–467
Huo L, Fan X, Xie Y, Yin J (2007) Short-term load forecasting based on the method of genetic programming. In: Proceedings of the 2007 IEEE international conference on mechatronics and automation, ICMA 2007, pp 839–843
Icke I, Bongard JC (2013) Improving genetic programming based symbolic regression using deterministic machine learning. In: 2013 IEEE congress on evolutionary computation, CEC 2013, pp 1763–1770
Koza JR (1992) Genetic programming: on the programming of computers by means of natual selection. Massachusetts
Koza JR (1994) Genetic programming as a means for programming computers by natural selection. Stat Comput 4(2):87–112
Lee HS, Seo WW, Lee KS (2019) Detection of Oak Wilt disease using convolutional neural network from Uav natural color imagery. In: IGARSS 2019 - 2019 IEEE international geoscience and remote sensing symposium. IEEE, pp 6622–6624
León Bañuelos LA (2019) Análisis de la distribución espacial de Arceuthobium Globosum implementando teledetección en el área de protección de flora y fauna Nevado de Toluca. PhD thesis, Universidad Autónoma del Estado de México
León-Bañuelos LA, Endara-Agramont AR, Gómez-Demetrio W, Martínez-García CG, Nava-Bernal EG (2020) Identification of Arceuthobium globosum using unmanned aerial vehicle images in a high mountain forest of central Mexico. J For Res 31(5):1759–1771
Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the international conference on computer vision, vol 2, pp 1150–1157
Luke S, Panait L (2020) Lexicographic parsimony pressure. In: GECCO 2002: proceedings of the genetic and evolutionary computation conference, pp 829–836
Mathiasen RL, Nickrent DL, Shaw DC, Watson DM (2008) Mistletoes: pathology, systematics, ecology, and management. Plant Disease 92(7):988–1006
MATLAB (2019) Image processing toolbox ^{TM}User ' s guide R 2019 b. Technical report, The MathWorks, Inc
Mihaylov R, Atanasow A, Ivanova A, Marinov A, Zahariev S (2020) Tracking the Development of Six Wheat Varieties Using Infrared Imaging and Image Processing Algorithms. *2020 International Conference Automatics and Informatics, ICAI 2020 - Proceedings*, 2020
Minařík R, Langhammer J, Lendzioch T (2020) Automatic tree crown extraction from UAS multispectral imagery for the detection of bark beetle disturbance in mixed forests. Remote Sens 12(24):1–31
Morales AK, Casas JG (2007) Algoritmos genéticos. Sociedad Mexicana de Inteligencia Artificial, Ciudad de México
Peñuelas J, Filella I (1995) Reflectance assessment of mite effects on apple trees. Int J Remote Sens 16(14):2727–2733

Peres DJ, Cancelliere A (2021) Analysis of multi-spectral images acquired by UAVs to monitor water stress of citrus orchards in sicily, Italy. In: World environmental and water resources congress, pp 270–278

Pernar R, Bajić M, Ančić M, Seletković A, Idžojtić M (2007) Detection of mistletoe in digital colour infrared images of infested fir trees. Period Biol 109:67–75

Poli R, Langdon WB, McPhee NF, Koza JR (2008) A field guide to genetic programing

Qi H, Wu Z, Zhang L, Li J, Zhou J, Jun Z, Zhu B (2021) Monitoring of peanut leaves chlorophyll content based on drone-based multispectral image feature extraction. Comput Electron Agric 187(June):106292

Rzedowski J, de Rzedowski GC (2011) Flora Del Bajío Y De Regiones Adyacentes. Instituto de Ecología 170:222–235

Sabrina F, Sohail S, Thakur S, Azad S, Wasimi S (2020) Use of deep learning approach on UAV imagery to detect mistletoe infestation. In: 2020 IEEE region 10 symposium, TENSYMP 2020, pp 556–559

Schmitt R (2002) 1 - introduction and survey of the electromagnetic. In: Electromagnetics explained. Newnes, pp 1–24

Shankar RLH, Veeraraghavan AK, Sivaraman K, Ramachandran SS (2018) Application of UAV for pest, weeds and disease detection using open computer vision. In: Proceedings of the international conference on smart systems and inventive technology, ICSSIT 2018, (Icssit), pp 287–292

Silva D, Almeida J (2007) GPLAB a genetic programming toolbox for MATLAB. In: Proceedings of the nordic MATLAB conference

Stehman SV (1996) Estimating the kappa coefficient and its variance under stratified random sampling. Photogram Eng Remote Sens 62(4):401–407

Thigpen J, Shah SK (2008) Multispectral imaging. In: Microscope image processing, 4th edn., pp 299–327

Universidad Nacional Autónoma de México (2022) Departamento de Botánica, Instituto de Biología (IBUNAM), Phoradendron velutinum (DC.) Oliv., ejemplar de: Herbario Nacional de México (MEXU), Plantas Vasculares

Walther D (2006) Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics. PhD thesis, California Institute of Technology

Xie Q, Huang W, Liang D, Chen P, Wu C, Yang G, Zhang J, Huang L, Zhang D (2014) Leaf area index estimation using vegetation indices derived from airborne hyperspectral images in winter wheat. IEEE J Sel Top Appl Earth Obs Remote Sens 7(8):3586–3594

Xue SY, Xu HY, Mu CC, Wu TH, Li WP, Zhang WX, Streletskaya I, Grebenets V, Sokratov S, Kizyakov A, Wu XD (2021) Changes in different land cover areas and NDVI values in northern latitudes from 1982 to 2015. Adv Climate Change Res

Yeom J, Jung J, Chang A, Maeda M, Landivar J (2017) Cotton growth modeling using unmanned aerial vehicle vegetation indices. In: IEEE international geoscience and remote sensing symposium (IGARSS), pp 5050–5052

Yuan Y, Fan W, Wang W, Liu H (2008) Robust collaborative optimization of a Multi-finger Micro-accelerometer based on genetic algorithm. In: Proceedings - 2nd international conference on genetic and evolutionary computing, WGEC 2008, vol 60474059, pp 105–108

Zhang N, Wang Y, Zhang X (2020) Extraction of tree crowns damaged by Dendrolimus tabulaeformis Tsai et Liu via spectral-spatial classification using UAV-based hyperspectral images. Plant Methods 16(1):1–20