Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5888

Crowd-Powered Sampling for Machine Learning: Leveraging Citizen Scientist Response Patterns in AutoML Workflows

Hugo Resende [**Institute of Science and Technology - Universidade Federal de São Paulo** | ***hresende@unifesp.br***]
Eduardo B. Neto [**Institute of Science and Technology - Universidade Federal de São Paulo** | ***ebneto@unifesp.br***]
Fabio A. M. Cappabianco [**Institute of Science and Technology - Universidade Federal de São Paulo** | ***cappabianco@unifesp.br***]
Álvaro L. Fazenda [**Institute of Science and Technology - Universidade Federal de São Paulo** | ***alvaro.fazenda@unifesp.br***]
Fabio A. Faria [**Instituto Superior Tecnico, Universidade de Lisboa** | ***fabio.faria@tecnico.ulisboa.pt***]

Institute of Science and Technology - Universidade Federal de São Paulo, Avenida Cesare Mansueto Giulio Lattes, n° 1201 - Eugênio de Mello, 12247-014, São José dos Campos, SP, Brazil.

Received: 11 April 2025 • **Accepted:** 24 September 2025 • **Published:** 16 March 2026

Abstract. Defining effective models for data classification is challenging, especially in complex contexts. Automated Machine Learning (AutoML) tools can assist in this process by generating rankings tailored to the nature of the data and the problem. In this work, we investigate the performance of five classifiers applied to the task of deforestation segment classification, using data labeled through a citizen science campaign from the ForestEyes project. We selected SVM, Ridge, AdaBoost, KNN, and MLP models based on a ranking generated with the PyCaret AutoML library, prioritizing diverse modeling approaches. Initially, the performance of the models is assessed using the incremental training strategy based on entropy of the volunteer's classifications. Then, a new training strategy is proposed based on the median response time of volunteers when evaluating each segment, exploring three ordering strategies: ascending, descending, and edge-based. Experimental results aligned with the PyCaret ranking, with SVM achieving the best performance, followed by Ridge and AdaBoost, especially when trained on smaller and more reliable data subsets. Both the entropy-based approach and the new strategy using median response time demonstrated strong potential to efficiently train machine learning models in scenarios with scarce data, typical in citizen science campaigns.

Keywords: Sampling Approaches, Citizen Science Data, AutoML, ForestEyes Project, Deforestation Detection

1 Introduction

In recent years, machine learning approaches have become the standard as a powerful tool for solving complex problems in various fields, ranging from healthcare, disease detection, to industry, and equipment failure prediction. However, selecting the best classifier for a given problem is not trivial and depends on several factors, such as the nature of the data and the evaluation metrics employed. In this context, Automated Machine Learning (AutoML) emerges as a promising approach to facilitate the search for the most suitable model. In particular, volunteer-labeled classification problems present additional challenges, as the quality of the labels may vary and influence the performance of machine learning models He *et al*. [2021]; Barbudo *et al*. [2023].

For such labeled data, citizen science initiatives (e.g., dedicated campaigns) are widely employed. In this regard, projects such as ForestEyes combine citizen science and machine learning for tropical forest monitoring [Dallaqua *et al*., 2021, 2022, 2019]. In this initiative, volunteers analyze small regions (segments) of remote sensing images from deforested areas and classify them as forest or deforestation. These labels are then used to train models that aim to automatically detect new deforested areas, contributing to environmental monitoring and conservation efforts [Fazenda and Faria, 2024].

Currently, the Support Vector Machine (SVM) classifier has been employed for deforestation detection within the ForestEyes project. However, no in-depth study has been conducted to assess whether other classifiers could produce better results, particularly in terms of balanced accuracy. In this sense, AutoML can be a valuable tool for identifying a more suitable classifier for the problem or confirming whether SVM remains the best option.

Previous research within the ForestEyes project has already explored different machine learning strategies for deforestation detection in remote sensing images. In particular, the study by Resende *et al*. [2024] proposed a training approach based on the entropy of volunteer responses, computed from the variability in the selection of majority labels for each segment. Haralick texture descriptors were sorted in ascending and descending orders of entropy, and an SVM model was progressively trained with different sample sizes, increasing by 5% at each step. In addition to this ordered approach, other strategies were tested, including edge-based selection (2.5% from each end, i.e., samples with the highest and lowest entropy), and a random selection approach. The results demon-

strated that it is possible to train the SVM with only 20% of the most reliable samples (lowest entropy) and achieve performance equivalent to training with the entire dataset. This finding is significant because citizen science data is valuable and scarce for requiring volunteer manual labeling.

In citizen science campaigns data labeling, we can not only utilize variability measures related to the content of the responses, such as Shannon entropy, but also analyze other variables that offer valuable insights, such as the response times of volunteers. For each segment, a predefined number of responses is expected in order to determine the most selected answer. In this context, labeling tasks completed with excessive haste leads to inaccuracies due to a lack of thorough review by the volunteer. Conversely, overly prolonged analyses may hinder efficiency and could reflect indecision or lack of confidence in the labeling process. By computing the median response time for each set of responses per segment, it becomes possible to more reliably investigate the influence of response time on the incremental training process of machine learning models, since the median is less sentitive to outliers.

This paper proposes a **novel training approach, based on the median response time of the volunteers** to assess the campain segments building upon the work proposed by Resende *et al*. [2024]. It also evaluates the performance of five classifiers trained on these segments. Classifier selection was guided by a ranking generated using the PyCaret AutoML library, applied the complete training dataset as in Resende *et al*. [2024], and was based on descriptive measures. The five classifiers chosen represent diverse solution-building strategies and do not necessarily correspond to the top five models in PyCaret's validation ranking.

Experimental results indicated that SVM stands out as the best-performing classifier for the classification of remote sensing data with specific characteristics. However, AdaBoost and Ridge classifiers also showed good performances, particularly when trained on smaller subsets of more reliable data—with low entropy or low median response times.

2 Theoretical Basis

In this section, we will detail the concepts related to the development of this research. Specifically, we will cover information about the Sentinel-2 satellite, the PRODES monitoring project, the SLIC and MaskSLIC segmentation algorithms, the PyCaret AutoML library, the Machine Learning methods used in this study, the Shannon Entropy and Homogeneity Rate metrics, and finally, the Haralick texture descriptors.

2.1 Sentinel-2 Satellite

The Sentinel-2 satellite was launched in 2015 by the European Space Agency (ESA) as part of a program known as Copernicus, which aims to monitor the planet and provide data for scientific research and sustainable public policies. Due to its spatial resolution, combined with its frequent revisit times (every 10 days, reduced to 5 days with the combined operation of the Sentinel-2A and Sentinel-2B satellites), Sentinel-2 is widely used for monitoring changes in land cover and vegetation dynamics, making it an important tool for studying environmental phenomena [Drusch *et al*., 2012; Main-Knorn *et al*., 2017].

Unlike other renowned satellites, such as Landsat-8, which has two instruments, Sentinel-2 is equipped with a single instrument, called the Multispectral Imager (MSI), designed to capture images in 13 spectral bands, covering wavelengths from the visible spectrum to the shortwave infrared. Since it has 10-meter resolution bands (blue, green, red, near-infrared, and SWIR) and 20-meter resolution bands (red edge), it allows the identification of surface details such as land use patterns, changes in vegetation health, and hydrological characteristics. Table 1 presents the main characteristics of the bands available in Sentinel-2 data.

Table 1. Spectral bands of the Sentinel-2 satellite.

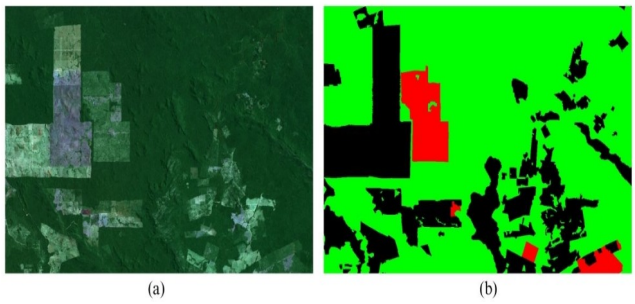
Band	**Central Wavelength (μm)**	**Spatial Resolution**
Band 1 (Coastal Aerosol)	0.443	60
Band 2 (Blue)	0.490	10
Band 3 (Green)	0.560	10
Band 4 (Red)	0.665	10
Band 5 (Red Edge 1)	0.705	20
Band 6 (Red Edge 2)	0.740	20
Band 7 (Red Edge 3)	0.783	20
Band 8 (Near Infrared)	0.842	10
Band 8A (Near Infrared)	0.865	20
Band 9 (Water Vapor)	0.945	60
Band 10 (Shortwave Infrared)	1.375	60
Band 11 (Shortwave Infrared)	1.610	20
Band 12 (Shortwave Infrared)	2.190	20

2.2 PRODES

The PRODES project is an initiative of the National Institute for Space Research (INPE), launched in 1988, aimed at monitoring and quantifying deforestation in the Legal Amazon [INPE, 2024]. It has become the primary reference for tracking forest cover loss, providing data that support public policies and conservation strategies. PRODES annually estimates the deforestation rate for the period from August to July [Gomes *et al*., 2014].

Initially, restricted to government use, the project was modernized with the launch of PRODES Digital in 2003, which made detailed digital maps publicly accessible [Valeriano *et al*., 2004]. Currently, its multi-temporal database, containing satellite images since 1988, enables the analysis of spatial and temporal deforestation patterns and the assessment of environmental policies. PRODES utilizes high-resolution optical sensors, such as Landsat-8, Sentinel-2 [Drusch *et al*., 2012], and CBERS-4 [Epiphanio, 2011], allowing differentiation between various land covers, including preserved vegetation and deforested areas.

Beyond monitoring, PRODES contributes to the development of deforestation assessment methodologies and serves as a reference (*ground-truth*) for validating image segmentation. Figure 1a illustrates one of the study areas analyzed in this work, while Figure 1b presents its corresponding *ground-truth*. In PRODES classification, red pixels indicate recent deforestation, green pixels correspond to preserved vegetation, and black pixels represent areas not analyzed, such as consolidated deforestation, water bodies, and infrastructure.



(a) (b)

Figure 1. Example of study area and its respective truth (*ground-truth*) PRODES.

2.3 SLIC and MaskSLIC Algorithms

The Simple Linear Iterative Clustering (SLIC) algorithm is a technique for segmenting color images by generating superpixels using the K-means clustering algorithm. It employs the CIELAB color space and takes as input parameters the desired number of superpixels (k) and a compactness factor (m) to ensure uniformity in the size and shape of superpixels. A key feature is its linear computational complexity relative to the number of pixels (N), as its search is limited to the predefined superpixel size [Achanta *et al.*, 2012].

Initially, the image is converted to the CIELAB color space, and k superpixels of size $\sqrt{N/k}$ pixels are created. Each superpixel is represented by a centroid $C_i = [L_i\ a_i\ b_i\ x_i\ y_i]^T$, where L, a, and b correspond to the average luminance and chromaticity values in the CIELAB color space, and x_i, y_i are the coordinates of the i-th superpixel. In each iteration, pixels are assigned to the nearest centroid within a limited region, and the centroids are updated with the average coordinates of their respective pixels. This process is repeated for a fixed number of iterations.

The distance measure D in SLIC combines color difference (d_c) and spatial distance (d_s) between pixels to determine their association with the nearest centroid, as presented in Equations 1, 2, and 3.

$$d_c = \sqrt{(L_i - L_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2} \quad (1)$$

$$d_s = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (2)$$

$$D = \sqrt{(d_c)^2 + \left(\frac{d_s}{S}\right)^2 \times m^2} \quad (3)$$

The MaskSLIC algorithm is an extension of SLIC designed to address challenges in superpixel generation, particularly in regions constrained by masks or regions of interest (RoIs), improving the precision and consistency of superpixel boundaries [Irving, 2016]. Unlike SLIC, which distributes seed points uniformly across the entire image, MaskSLIC spatially distributes seed points only within the designated mask by applying an Euclidean distance transform to ensure strategic coverage of the RoI.

2.4 PyCaret AutoML Tool

The increasing demand for efficient and accessible machine learning models has driven the development of AutoML approaches. AutoML aims to automate the most complex steps of model development, such as algorithm selection, hyperparameter tuning, feature engineering, and cross-validation. This approach democratizes machine learning by allowing both experts and non-experts to build predictive models without requiring deep technical knowledge. Additionally, AutoML reduces experimentation time and improves the reproducibility of results, making it a valuable solution in scenarios where speed and efficiency are crucial [He *et al.*, 2021; Barbudo *et al.*, 2023].

In this sense, PyCaret is a Python-based AutoML library that simplifies the process of building and deploying machine learning models. It provides a unified interface that enables users to train, evaluate, and compare multiple algorithms with just a few lines of code. Its key features include automated data preprocessing, automatic selection of the best model based on performance metrics, and easy deployment. Furthermore, PyCaret supports both supervised and unsupervised learning, offering modules for classification, regression, clustering, and dimensionality reduction [Ali, 2020].

Beyond its simplicity and efficiency, PyCaret also stands out for its integration with various popular data science tools such as Pandas, Scikit-learn, and MLflow. This compatibility allows users to seamlessly incorporate PyCaret into existing workflows, accelerating model development and experimentation. As a result, PyCaret represents an accessible and powerful solution for implementing AutoML in a wide range of applications.

2.5 Shannon's Entropy

Shannon entropy is a widely used metric for measuring uncertainty and information dispersion in a dataset. Its application spans multiple fields, including computer science, statistics, information theory, and physics, where it is employed in problems ranging from data compression to analyzing the complexity of dynamic systems [Lin, 1991]. Mathematically, this metric is defined by Equation 4.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2(p(x_i)) \quad (4)$$

In Equation 4, $p(x_i)$ denotes the probability of occurrence of event x_i, and n is the total number of classes. In the context of this study, this probability is associated with the frequency at which a given class (*forest* or *non-forest*) is selected by volunteers. Entropy is calculated as the ratio between the

number of votes assigned to the i-th class and the total number of votes collected for the corresponding segment (or task).

In this research, entropy plays a central role in assessing the complexity of the analyzed tasks. High entropy values indicate greater dispersion in volunteer responses, revealing a higher level of uncertainty and classification difficulty. Thus, in addition to quantifying response variability, entropy also provides an objective estimate of the level of challenge posed by each task within the citizen science campaign.

2.6 Homogeneity Rate

he quality of a segmentation can be assessed using various metrics, with the Homogeneity Rate (HoR) being one of the most relevant for measuring the uniformity within a segment. This metric is computed based on the proportion of pixels belonging to the predominant class in a binary segment and was originally proposed to determine whether a region was mainly composed of forest or deforestation pixels [Dallaqua *et al*., 2021]. HoR values close to 1 indicate more homogeneous and easily interpretable segments for volunteers, whereas values near 0.5 suggest a balanced mixture of classes, making classification more challenging and increasing uncertainty about the actual composition of the region.

The mathematical formulation of HoR is presented in Equation 5, where NFP and NNP represent the number of pixels belonging to the forest and non-forest classes, respectively, while NP corresponds to the total number of pixels in the segment.

$$HoR = \frac{\max(\text{NFP}, \text{NNP})}{\text{NP}} \tag{5}$$

2.7 Haralick Texture Features

The Haralick texture descriptors refer to a set of 14 features used to represent the texture of an image based on the gray-level co-occurrence matrix (GLCM). The GLCM records the frequency at which pairs of pixels with specific values occur together, considering a certain distance and direction. To cover different orientations in the image, four matrices corresponding to directions of 0, 45, 90, and 135 degrees are computed and then normalized to form a probability matrix. From this matrix, Haralick's 14 descriptors are extracted [Haralick *et al*., 1973].

Among these descriptors, (i) the angular second moment expresses GLCM uniformity, indicating homogeneous textures; (ii) contrast measures the intensity variation between a pixel and its neighbors, highlighting textural differences; (iii) correlation quantifies the linear relationship between adjacent pixels, reflecting their intensity dependency; (iv) sum of squares (variance) describes the dispersion of values around the mean, associated with intensity diversity in the image; (v) inverse difference moment evaluates texture uniformity, assigning higher values to elements near the GLCM diagonal; (vi) sum average calculates the weighted mean of summed pixel intensities, indicating their central tendency; (vii) sum variance measures the variability of summed intensities, pointing to the diversity of textural patterns; (viii) sum entropy quantifies the degree of randomness in summed pixel intensities; (ix) entropy measures the complexity or disorder of the image texture; (x) difference variance assesses the dispersion of intensity differences within the GLCM; (xi) difference entropy measures the degree of randomness in intensity differences between adjacent pixels; (xii) and (xiii) two correlation information metrics quantify the mutual dependence between gray levels, revealing intensity patterns in the image; and (xiv) maximal correlation coefficient determines the highest correlation between pixel intensities, reflecting similarity across different regions of the image.

2.8 Machine Learning Classifiers

In this subsection, the definitions and main characteristics of the classifiers used in this research will be presented.

2.8.1 Ridge

The Ridge Classifier is an adaptation of Ridge Regression, which is a linear regression technique with regularization to reduce multicollinearity and prevent overfitting. Multicollinearity complicates the analysis of correlated predictor variables, while overfitting occurs when the model captures noise in the training data. Ridge Regression addresses this with a regularization term [Schreiber-Gregory, 2018].

The Ridge Classifier adapts Ridge Regression for classification tasks, assigning samples to predefined classes instead of predicting continuous values. The model uses a cost function with squared error and L2 regularization, helping to deal with multicollinearity. This approach is efficient for binary and multiclass classification problems, especially with collinear data, and is computationally more efficient than Logistic Regression in some cases [Peng and Cheng, 2020; Saunders *et al*., 1998].

2.8.2 AdaBoost

Adaptive Boosting (AdaBoost) is a machine learning method that improves the accuracy of weak classifiers by combining them iteratively. Instead of training a single model, AdaBoost builds a sequence of classifiers by adjusting their weights to focus on hard-to-classify examples [CAO *et al*., 2013; Liu *et al*., 2024].

The process begins with a weak classifier, such as a simple decision tree. After training, misclassified samples are given higher weights, and a new classifier is trained with these updated weights. This process is repeated until a predefined number of classifiers is reached or an acceptable error level is achieved. The final prediction is made through a weighted vote among the classifiers.

AdaBoost enhances accuracy without extensive tuning and resists overfitting when using simple classifiers.

However, its performance can be affected by noisy data or overly complex weak classifiers, which may lead to overfitting.

2.8.3 Support Vector Machine

The Support Vector Machine (SVM) is a classifier used for classification and regression, aiming to find an optimal hyper-

plane that maximizes class separation. SVM training involves finding this hyperplane to maximize the margin between it and the support vectors. For linearly separable data, SVM uses a straight hyperplane. When linear separation is not possible, a technique known as the kernel trick is applied, projecting the data into a higher-dimensional space. The most common kernels include linear, polynomial, Gaussian (RBF), and sigmoid [Roy and Chakraborty, 2023].

One of the main advantages of SVM is its robustness against overfitting, especially in high-dimensional datasets. Additionally, it can model nonlinear relationships using kernels. However, its training can be computationally expensive for large datasets, and the choice of the appropriate kernel can significantly impact performance.

2.8.4 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) is a supervised algorithm that classifies data based on similarity with its nearest neighbors. It calculates the distance between a given data point and all examples in the training set, selects the k closest neighbors, and assigns the most frequent class (classification) or the average of the values (regression). The Euclidean distance is one of the most commonly used metrics for this task [Halder *et al*., 2024; Uddin *et al*., 2022].

Among the main advantages of KNN are its simplicity and the absence of explicit training, as all operations occur during the prediction phase. However, it also presents challenges, such as high computational cost for large datasets and the sensitivity of the k value, which can impact model accuracy. Additionally, data normalization can be a crucial factor in preventing distortions in proximity measurements [Jodas *et al*., 2022].

2.8.5 Multi-layer Perceptron

The Multilayer Perceptron Neural Network (MLP) is a learning model based on artificial neural networks, used for classification tasks. Unlike linear classifiers, MLP employs multiple layers of neurons to learn nonlinear data representations. The training process relies on specific algorithms such as backpropagation, which adjusts the weights of neural connections using gradient descent to minimize the error function. To prevent overfitting, techniques such as L2 regularization, dropout, and weight normalization can be applied [Wu and Feng, 2018; Tashakkori *et al*., 2024].

Typically, MLP uses nonlinear activation functions, such as ReLU in hidden layers and softmax in the output layer, for multiclass classification. Its main advantages include high generalization capability and flexibility for various applications. However, deeper networks may require high computational power and careful hyperparameter tuning.

3 ForestEyes Project

The ForestEyes project, by combining citizen science and machine learning in tropical forest monitoring, plays a highly significant role. Due to the complexity of its multiple stages, an activity diagram is necessary to properly organize its workflow. In this regard, to better structure the project, five modules have been defined, as illustrated in Figure 2. These modules will be further detailed in the following subsections.

3.1 Preprocessing

he Preprocessing module of the ForestEyes project consists of three main stages: acquisition, processing, and segmentation. In the acquisition stage, satellite images from sources such as Landsat-8 and Sentinel-2 are obtained, ensuring comprehensive coverage of the areas of interest. Next, in the processing stage, Geographic Information System (GIS) tools, such as QGIS, are used to prepare the data [QGIS, 2025]. During this phase, operations such as cropping the areas of interest and aligning the images with PRODES reference data are performed. Finally, in the segmentation stage, the processed images are divided into segments, ignoring certain regions, such as water bodies and areas of consolidated deforestation, while prioritizing recently deforested areas, as identified in the latest PRODES report.

Specifically in this research, after selecting the geographic region of interest, Sentinel-2 images containing bands B1 (coastal aerosol), B2 (blue), B3 (green), B4 (red), B8 (VNIR), B11 (SWIR), and B12 (SWIR) were collected. In the context of this investigation, nine areas located in the Xingu River Basin, in the state of Pará, were analyzed, covering a total of 8,514 hectares. After collecting the bands, a composition of three of them was performed before the segmentation process. Based on preliminary experiments using the SLIC and MaskSLIC algorithms, it was identified that the composition of bands B4, B3, and B2 (corresponding to the RGB spectrum) provides higher-quality segmentations. Thus, after this composition, the MaskSLIC algorithm was applied, allowing irrelevant regions, such as areas of consolidated deforestation, to be ignored during the segmentation process.

3.2 Citizen Science

The Citizen Science module is structured into three main stages: task construction, campaign construction on the citizen science platform, and volunteer responses. The task construction phase is based on the segments obtained in the segmentation module, where each image is assigned a ”segmentation map” that attributes a numerical value to each pixel, corresponding to the segment it belongs to. This map allows for the individual extraction of segments and their projection onto different false-color compositions selected for the campaign. In the next phase, the campaign is designed and structured within the Citizen Science platform, where workflow parameters are configured, task sets are defined, and explanatory materials are prepared to guide participants. The workflow determines the sequence of user interactions, from image presentation to classification submission. A critical setup aspect is determining the minimum responses per task to ensure reliable majority classification. This decision directly impacts the robustness of the analysis and can be adjusted based on image complexity and the expected error rate. To organize image presentation, they are grouped into logical sets (subject sets) within the Zooniverse platform. Once this

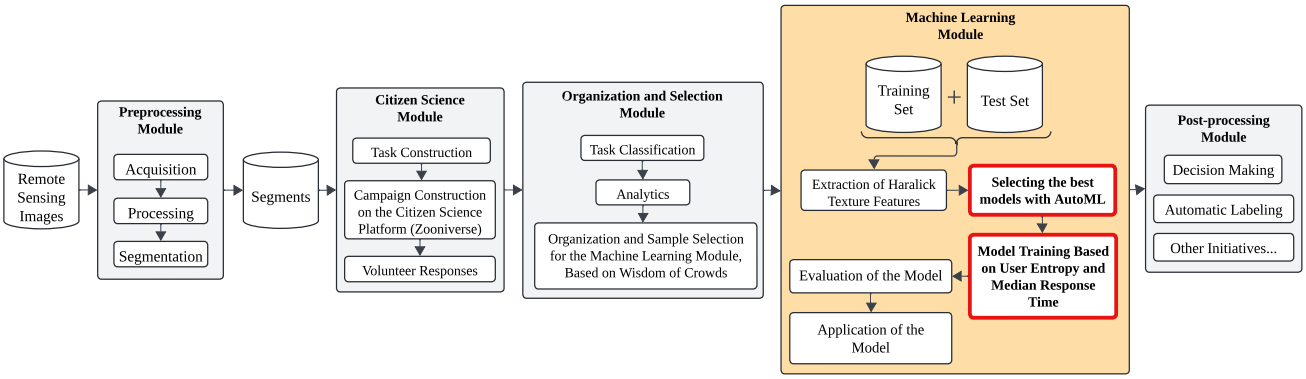


Figure 2. ForestEyes project's schematic representation. The orange module correspond to the module implemented in this work.

structuring is complete, the workflow is published and made available to volunteers, who actively contribute to the image classification process.

Similar to the study by Resende *et al*. [2024], this research utilized different false-color compositions to represent each segment (task), including the combinations B8B11B4, B4B11B12, and B11B8B4, along with the NDVI. For the campaign setup, 90 segments with HoR = 1 and 90 segments with HoR between 0.7 and 0.8 were selected, allowing for an assessment of how homogeneity influences volunteer decision-making. After defining the tasks, the campaign was configured on the Zooniverse platform, establishing the workflow and participation criteria. Over approximately two weeks, volunteer responses were collected and analyzed, enabling a detailed investigation of how segment variability affects classification accuracy. This process provided insights into how different false-color compositions and NDVI aid in the visual interpretation of images, contributing to improving the effectiveness of the Citizen Science approach.

3.3 Organization and Selection

In the sample organization and selection module, after the campaign is completed, the volunteers' responses are processed and analyzed. Specifically, this module handles data that provide insights into the campaign, including the response time for each task and the variability of volunteers' responses in determining the majority classification for each segment (task). In this research, variability is assessed through the entropy calculation for each task. Similar to the study by Resende *et al*. [2024], a ranking of tasks based on entropy was constructed. Based on this ranking, the segments associated with each task were selected to form the training set for the machine learning module. This strategy enabled a detailed analysis of the reliability of volunteer classifications and helped in selecting the most suitable samples for training the top-ranked machine learning models in the experimentation with the AutoML library.

3.4 Machine Learning

In the Machine Learning Module, the training and test sets are constructed by computing the Haralick texture descriptors for each segment generated in the previous stages. Differently from the study by Resende *et al*. [2024], which directly applied the SVM technique with a linear kernel—following approaches previously explored within the ForestEyes project—this research introduces a preliminary step of automated model selection. This selection is carried out through an AutoML strategy, which evaluates various classification algorithms trained on the entire training dataset and identifies the top five performers on validation stage (out of the top-10), ensuring a diversity of modeling approaches. Particularly, the SVM (#1), Ridge (#2), AdaBoost (#3), KNN (#7) and MLP (#10) methods were investigated.

The model selection is guided by performance metrics such as F1-score, balanced accuracy, recall, and other complementary indicators, all computed during validation, in order to ensure robustness and generalization of the results. Based on this process, the best-performing classifiers are selected for application and further analysis in new areas of interest. This approach allows greater flexibility in comparing algorithms and provides a more rigorous assessment of how sample variability influences model performance.

In addition to the AutoML-based selection of classifiers, this work also adopts specific sampling strategies for the construction of training sets, aiming to investigate the impact of variability in volunteer classifications. One of these strategies is based on the entropy of responses for each segment (task), as proposed by Resende *et al*. [2024]. From this, several training sets were created incrementally, starting with the 5% of segments with the lowest and highest entropy and progressively adding 5% more until reaching 100% of the available samples. Furthermore, the 'edges' approach was explored incrementally, selecting 2.5% of the samples with the highest and 2.5% with the lowest entropy values, along with a random sampling strategy for comparative analysis.

Additionally, this research proposes a **new incremental sampling approach based on the median response times of volunteers for each task**. The use of the median, rather than the mean, for example, aims to mitigate the influence of outliers caused by extremely fast or slow responses, which could significantly distort the analysis. Similar to the entropy-based strategy, the segments were ordered based on median response times and included in training sets in 5% increments. This allowed the construction of performance curves for the models as a function of the inferred reliability from user response times. This novel strategy contributes to a deeper understanding of the relationship between volunteer engage-

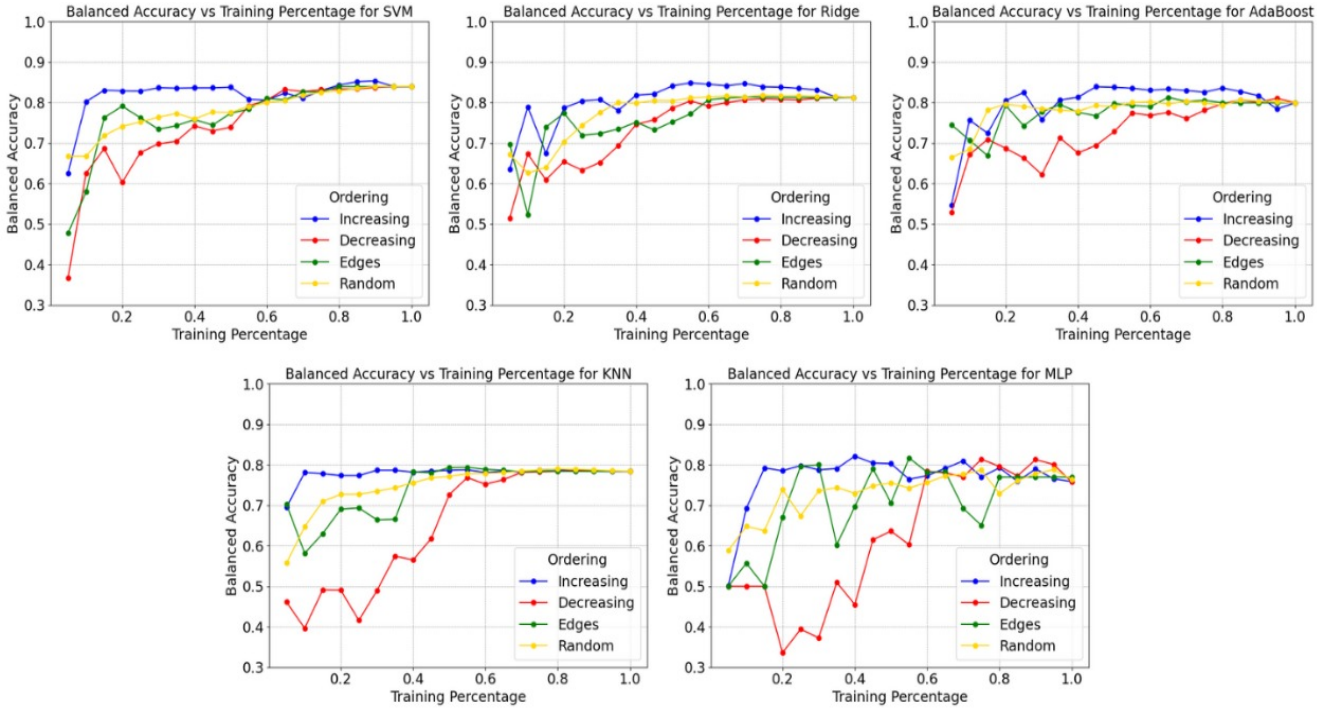


Figure 3. Performance of classifiers trained with the **entropy-based approach** [Resende *et al*., 2024].

ment and the quality of data generated through citizen science campaigns.

Regarding the sample sizes, and following the methodology of Resende *et al*. [2024], 96 segments from the 'forest' class and 83 from the 'non-forest' class—collected through the Sentinel-2 campaign—were used in the training set. The test set comprised 62,643 segments in total, including 54,771 labeled as 'forest' and 7,872 as 'non-forest'. It is important to note that for each segment, 13 Haralick features were extracted in four different directions, yielding a substantial amount of numerical data to support robust training of machine learning models.

3.5 Post processing

In the post-processing module, the final decision-making steps are carried out based on the automatic classifications generated by the models selected in the machine learning stage. In this phase, the results produced by the classifiers are organized and interpreted to enable the identification of regions with a higher probability of deforestation, according to the distinction between segments classified as "forest" or "non-forest."

Moreover, this final stage also opens up opportunities for other initiatives within the ForestEyes project, such as the automatic labeling of new images based on previously trained models. From the most reliable results, it is possible to continuously feed the system with newly labeled samples, supporting both continuous monitoring and the construction of expanded datasets. Other possibilities include the automatic suggestion of regions of interest for future citizen science campaigns, directing volunteer participation toward more uncertain or critical areas, contributing to an iterative cycle of improvement in both data quality and model effectiveness.

4 Experimental Results and Discussion

This section presents the experimental results obtained from training the top-performing models, based on distinct approaches, selected through the PyCaret AutoML library. Subsection 4.1 introduces the results of the entropy-based approach, as proposed in the work of Resende *et al*. [2024], considering the performance of these models. Next, Subsection 4.2 discusses the outcomes associated with the new training approach introduced in this study, which is based on the median response times recorded for each task. Finally, in Subsection 4.3, the relationship between the results of both approaches is analyzed, highlighting similarities, differences, and implications for future applications

4.1 Entropy-based Approach

The strategy based on increasing entropy ordering - that is, starting with the most reliable samples, which exhibit the lowest variability in volunteer responses - showed very promising results (see Figure 3). As previously discussed by Resende *et al*. [2024], the SVM classifier achieved balanced accuracy levels close to those obtained with the full training set, even when trained with only the most reliable 10% of the samples. This pattern was also observed for the other classifiers, except for the MLP, which required 15% of the most reliable samples to reach similar performance. Overall, progressively adding samples with higher entropy (and therefore lower confidence) did not bring significant improvements to the classifiers; in fact, the top 20% highest-entropy samples often harmed model performance. In almost all classifiers, except for the SVM, models trained with 80% of the samples outperformed those trained with the full dataset, indicating

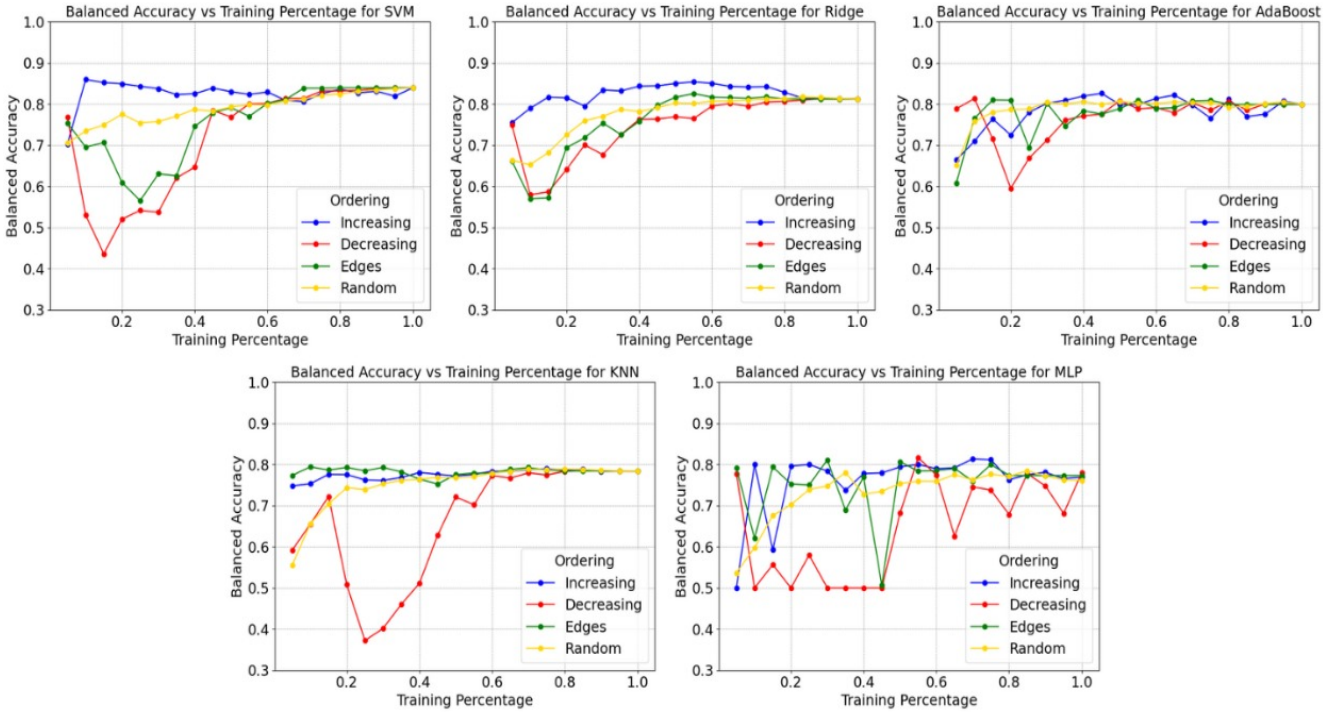


Figure 4. Performance of classifiers trained with the **median response time-based approach**.

that excluding the least reliable samples can even enhance overall accuracy.

In contrast, the decreasing entropy strategy — which prioritizes samples with high entropy — did not yield satisfactory results. In general, classifiers only began to stabilize their performance when the training set included at least 60% of the samples, meaning a considerable portion of more reliable data had to be incorporated. This highlights the fragility of this strategy when relying mostly on low-confidence samples. The MLP case was particularly sensitive: when trained with 20% to 55% of the least reliable samples, the model exhibited erratic behavior, with no clear trend of accuracy improvement as more data was added. These findings suggest that this strategy is not only inefficient but may also lead to unstable models, making it ill-suited for the ForestEyes context.

The border strategy - which combines samples with both high and low entropy in symmetric proportions - proved superior to the decreasing strategy, but still inferior to the increasing one. While it produced reasonable results in some scenarios, it generally failed to outperform models trained with only the most reliable samples. A notable exception was the MLP, which slightly surpassed the increasing strategy when trained with 55% of the samples. However, this was a very specific case that does not undermine the broader evidence supporting the increasing entropy approach, especially considering its efficiency in requiring fewer training samples while consistently delivering more robust performance across all classifiers.

When comparing classifiers under the increasing entropy strategy, SVM, KNN, and MLP were the models that reached stable performance with the fewest reliable samples — specifically, 15%, 10%, and 15%, respectively. Nevertheless, the SVM stood out as the classifier with the best balanced accuracy across nearly all scenarios, regardless of the ordering strategy used. Therefore, SVM emerges as the most suitable model for the ForestEyes project, particularly due to its ability to perform well with smaller, yet highly reliable training sets, in the context of remote sensing image classification.

4.2 Median Response Time-based Approach

The strategy based on the ascending ordering of median response times (i.e., using first the samples derived from tasks that required less time from volunteers) yielded the best results among all strategies analyzed based on this dataset (see Figure 4). Specifically, the SVM trained with only 10% of the samples achieved the highest balanced accuracy among all combinations of approaches and classifiers. In general, except for the MLP trained with up to 20% of the samples with the lowest median response times, the other classifiers exhibited very little fluctuation in balanced accuracy regardless of the training set size. This indicates that samples derived from tasks analyzed more quickly (lower median response time) lead to more efficient training, including excellent model convergence.

In the descending strategy, which prioritized samples from tasks with higher median response times, it was observed that all classifiers, without exception, achieved good balanced accuracy using only 5% of the samples. However, as the percentage of samples increased from 5% to 20%, there was noticeable instability in the learning process, with fluctuations in model performance. This strategy, therefore, supports the findings from the ascending approach, suggesting that tasks analyzed more quickly (noting that we are considering the median of 15 responses) tend to provide more reliable data, whereas tasks that required more time may reflect greater uncertainty among volunteers. This factor, in turn, appears to negatively impact classifier performance.

The edge-based strategy, which combines samples with

both the lowest and highest median response times, showed varied results depending on the classifier. For the SVM and Ridge Classifier, the behavior was very similar to that observed in the descending strategy. On the other hand, for AdaBoost and especially KNN, this strategy outperformed the descending one, even surpassing the ascending strategy in the case of KNN. The MLP, in turn, only exhibited a stable learning curve when trained with 45% of the samples (22.5% with the lowest and 22.5% with the highest median response times), which highlights unstable behavior with smaller datasets. Due to these classifier-specific responses, the edge-based strategy did not prove to be a consistent or viable training strategy for any of the classifiers recommended by PyCaret. For this reason, it is not considered the most suitable for applications such as the ForestEyes project.

When comparing classifier performance based on the median response time ordering approach, the SVM consistently outperformed the other models across nearly all scenarios. Both in the ascending strategy with only 10% of the samples and with the full training set, SVM achieved the highest balanced accuracy. In contrast, the MLP exhibited significant instability and, overall, produced the worst results. The analysis of performance curves further suggests that, in the case of SVM, including more reliable samples could potentially lead to additional gains in performance, while the other models appear to have already reached their performance plateau with the current amount of training data.

4.3 Analysis Between Entropy-based and Response Time-based Approaches

The two analyzed approaches - one based on response entropy and the other on the median response time per task - proved to be quite interesting and complementary. The first considers the variability in the content of responses provided by volunteers, quantified through Shannon entropy. The second is based on a factor that is indirectly related to the content, related to the median time volunteers took to analyze each task. Despite the conceptual differences between these approaches, a notable similarity was observed in the performance patterns of the classifiers across the increasing, decreasing, and edge-based strategies.

In general, the entropy-based training approach with the increasing strategy yielded good results for almost all classifiers. The exceptions were AdaBoost and MLP, especially when trained with small sample sizes. Interestingly, this trend was also observed in the approach based on median response times. Classifiers such as SVM, Ridge, and KNN achieved high balanced accuracy even with reduced training sets, as long as the samples were associated with tasks that had shorter median response times. The fact that this trend was confirmed in both approaches is a highly relevant finding, as it demonstrates that effective models can be trained with a fraction of the samples, provided that these samples are selected based on quality criteria such as entropy or response time.

On the other hand, the decreasing strategy, both in the entropy-based and the median response time-based approaches, did not yield satisfactory results. In both cases, it was necessary to use at least 50% of the training samples for the models to achieve more stable balanced accuracy values. This indicates that samples derived from tasks with high entropy or longer response times may introduce uncertainty and noise into the model training process.

Finally, it is important to emphasize that the MLP did not perform well in either of the analyzed approaches. This classifier showed instability and low balanced accuracy values, even in scenarios where the other models produced satisfactory results. This suggests that the MLP, at least in the automated configuration used in this study, is not the most suitable choice for this type of problem, being consistently outperformed by classifiers such as SVM, Ridge, and KNN.

5 Conclusion

In this study, the behavior of five classifiers was investigated in the context of classifying deforestation segments, based on data labeled through a citizen science campaign and using the approach presented by Resende *et al*. [2024]. The selection of classifiers was guided by a ranking generated with the support of the AutoML library PyCaret during the validation stage. To explore different solution construction strategies, five classifiers were chosen from among the top ten ranked models, prioritizing diversity in their underlying approaches.

With the classifiers SVM, Ridge, AdaBoost, KNN, and MLP defined, the incremental training approach proposed by Resende *et al*. [2024] was explored, and the performance of each model was evaluated. Additionally, this study proposed a novel training approach based on a behavioral factor that is indirectly related to the content of the responses: the median response time for each segment. The samples were ordered based on their median response times in ascending and descending order, as well as using the edge-based strategy, similarly to the methodology of Resende *et al*. [2024]. The results showed that SVM achieved the best performance, with Ridge and AdaBoost also yielding strong results, especially when trained with small, high-confidence data subsets. Interestingly, the performance order of the classifiers matched the ranking produced by PyCaret, with KNN and MLP occupying the second-to-last and last positions, respectively. Regardless of the variations between classifiers, a consistent behavior was observed across both training approaches.

Thus, selecting classifiers based on rankings generated by AutoML tools such as PyCaret proved to be a viable and effective strategy, as the experimental results closely followed the expected performance order. Furthermore, both the entropy-based approach and the new approach proposed in this study—based on median response times—proved to be promising, particularly for enabling the training of machine learning models with small datasets without significantly compromising performance. This is especially relevant given the typically limited nature of available data in many citizen science campaigns, such as the ForestEyes project.

6 Future Works

As a natural continuation of this research, there is a clear need to evaluate the performance of the classifiers used, combined with the investigated approaches, on data from other regions

of the Brazilian Legal Amazon. This geographic expansion could help assess the robustness and generalization capacity of the models in different environmental and social contexts. Furthermore, it is suggested that these strategies be applied to data from other Brazilian biomes, such as the Cerrado and the Atlantic Forest, where land use and land cover dynamics also require ongoing monitoring and change detection efforts.

Another promising direction for future work is the evaluation of these classifiers and approaches using data from different remote sensing sensors. In particular, the use of Synthetic Aperture Radar (SAR) imagery stands out, as it has proven valuable for monitoring tropical forests due to its ability to capture data even under cloud cover. This research line is especially relevant for projects like ForestEyes, which already explores the use of both optical and SAR sensors in deforestation detection.

Declarations

Authors' Contributions

Hugo Resende contributed to data organization, part of the experimentation (AutoML), and was responsible for the organization and writing of this manuscript. Eduardo B. Neto conducted most of the experiments in this study (Haralick feature extraction, model training, and evaluation) and manuscript revision. Fabio A. M. Cappabianco contributed with supervision and manuscript revision. Álvaro L. Fazenda was responsible for funding acquisition, project supervision and management, and manuscript revision. Fabio A. Faria was the main creator of this research and also contributed to the monitoring of the computational experiments and manuscript revision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the UNIFESP and the IFSULDEMINAS for their support; the CAPES for financial assistance; the LNCC for providing the HPC computational resources and the Zooniverse platform. This research received partial funding from the São Paulo Research Foundation (FAPESP, Brazil) under grants #2023/00811-0, #2023/00782-0, and #2024/01115-0. The latter also supports activities of the Center of Science for the Development of Carbon Neutral Cities.

Availability of data and materials

The datasets (and/or softwares) generated and/or analyzed during the current study will be made upon request.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282. DOI: 10.1109/TPAMI.2012.120.

Ali, M. (2020). *PyCaret: An open source, low-code machine learning library in Python*. Available at:`https://www.pycaret.org` PyCaret version 1.0.

Barbudo, R., Ventura, S., and Romero, J. R. (2023). Eight years of automl: categorisation, review and trends. *Knowledge and Information Systems*, 65(12):5097–5149. DOI: 10.1007/s10115-023-01935-1.

CAO, Y., MIAO, Q.-G., LIU, J.-C., and GAO, L. (2013). Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, 39(6):745–758. DOI: 10.1016/S1874-1029(13)60052-X.

Dallaqua, F. B., Fazenda, Á. L., and Faria, F. A. (2021). Foresteyes project: Conception, enhancements, and challenges. *Future Generation Computer Systems*, 124:422–435. DOI: 10.1016/j.future.2021.06.002.

Dallaqua, F. B. J. R., Faria, F. A., and Fazenda, □L. (2022). Building data sets for rainforest deforestation detection through a citizen science project. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5. DOI: 10.1109/LGRS.2020.3032098.

Dallaqua, F. B. J. R., Fazenda, Á. L., and Faria, F. A. (2019). Foresteyes project: Can citizen scientists help rainforests? In *2019 15th International Conference on eScience (eScience)*, pages 18–27. IEEE. DOI: 10.1109/escience.2019.00010.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P. (2012). Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36. The Sentinel Missions - New Opportunities for Science. DOI: 10.1016/j.rse.2011.11.026.

Epiphanio, J. C. N. (2011). Cbers-3/4: características e potencialidades. In *Proceedings of the Brazilian Remote Sensing Symposium, Curitiba, Brazil*, volume 30, page 90099016. Available at:`https://www.semanticscholar.org/paper/CBERS-3-4%3A-caracter%C3%ADsticas-e-potencialidades-Epiph%C3%A2nio/8ac613c391f3396fff0295b7949a2728f93e9d24`.

Fazenda, A. L. and Faria, F. A. (2024). Foresteyes: Citizen scientists and machine learning-assisting rainforest conservation. *Communications of the ACM*, 67(8):95–96. DOI: 10.1145/3653319.

Gomes, A. R., Diniz, C. G., and Almeida, C. A. (2014). Amazon regional center (inpe/cra) actions for brazilian amazon forest: Terraclass and capacity building projects. *Interdiscip. Analysis and Modeling of Carbon-Optimized Land Manag. Strategies for Southern Amazonia*, page 101. Book.

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., and Khraisat, A. (2024). Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1):113. DOI: 10.1186/s40537-024-00973-y.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621. DOI: 10.1109/TSMC.1973.4309314.

He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the

state-of-the-art. *Knowledge-based systems*, 212:106622. DOI: 10.1016/j.knosys.2020.106622.

INPE (2024). PRODES - Project for Monitoring Deforestation in the Legal Amazon by Satellite. Available at:`https://www.obt.inpe.br/prodes/index.php` Acessado em setembro/2024.

Irving, B. (2016). maskslic: regional superpixel generation with application to local pathology characterisation in medical images. *arXiv preprint arXiv:1606.09518*. DOI: 10.48550/arxiv.1606.09518.

Jodas, D. S., Passos, L. A., Adeel, A., and Papa, J. P. (2022). Pl-k nn: A parameterless nearest neighbors classifier. In *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–4. IEEE. DOI: 10.1109/iwssip55020.2022.9854445.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151. DOI: 10.1109/18.61115.

Liu, B., Li, X., Xiao, Y., Sun, P., Zhao, S., Peng, T., Zheng, Z., and Huang, Y. (2024). Adaboost-based svdd for anomaly detection with dictionary learning. *Expert Systems with Applications*, 238:121770. DOI: 10.2139/ssrn.4379462.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., and Gascon, F. (2017). Sen2Cor for Sentinel-2. In Bruzzone, L., editor, *Image and Signal Processing for Remote Sensing XXIII*, volume 10427, page 1042704. International Society for Optics and Photonics, SPIE. DOI: 10.1117/12.2278218.

Peng, C. and Cheng, Q. (2020). Discriminative ridge machine: A classifier for high-dimensional data or imbalanced data. *IEEE transactions on neural networks and learning systems*, 32(6):2595–2609. DOI: 10.1109/tnnls.2020.3006877.

QGIS (2025). Qgis - geographic information system. Available at:`https://qgis.org`.

Resende, H., Neto, E. B., Cappabianco, F. A., Fazenda, A. L., and Faria, F. A. (2024). Sampling strategies based on wisdom of crowds for amazon deforestation detection. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. IEEE. DOI: 10.1109/sibgrapi62404.2024.10716332.

Roy, A. and Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*, 233:109126. DOI: 10.1016/j.ress.2023.109126.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. page 515–521. Available at:`https://dl.acm.org/doi/10.5555/645527.657464`.

Schreiber-Gregory, D. N. (2018). Ridge regression and multicollinearity: An in-depth review. *Model Assisted Statistics and Applications*, 13(4):359–365. DOI: 10.3233/mas-180446.

Tashakkori, A., Talebzadeh, M., Salboukh, F., and Deshmukh, L. (2024). Forecasting gold prices with mlp neural networks: A machine learning approach. *International Journal of Science and Engineering Applications (IJSEA)*, 13:13–20. DOI: 10.7753/ijsea1308.1003.

Uddin, S., Haque, I., Lu, H., Moni, M. A., and Gide, E. (2022). Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. *Scientific Reports*, 12(1):6256. DOI: 10.1038/s41598-022-10358-x.

Valeriano, D. M., Mello, E. M., Moreira, J. C., Shimabukuro, Y. E., Duarte, V., Souza, I., Santos, J., Barbosa, C. C., and Souza, R. (2004). Monitoring tropical forest from space: the prodes digital project. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 35:272–274. Available at:`https://www.isprs.org/proceedings/xxxv/congress/comm7/papers/53.pdf`.

Wu, Y.-c. and Feng, J.-w. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, 102:1645–1656. DOI: 10.1007/s11277-017-5224-x.